Facial Detection of Genetic Disorders

Mohsan Shahzad Alvi

Lady Margaret Hall
University of Oxford

Submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

Trinity 2019

Abstract

An estimated 400,000 children are born every year with rare genetic disorders that significantly affect their quality of life. Early detection and intervention can significantly improve the quality of life of these children. Craniofacial characteristics contain highly useful information for clinical geneticists for diagnosis. This thesis investigates the use of computer vision to aid in the automatic detection of genetic disorders from ordinary facial photographs. This is a non-trivial task, in part due to patient privacy concerns and the scarcity of training data. In the following, we present several approaches to overcome these challenges.

First, we present a method for creating realistic-looking average faces for individuals sharing a syndrome. These averages remove identifiable features, but retain clinically relevant phenotype information and preserve facial asymmetry. This procedure is completely automated, removing the need to expose patient identities at any point during the process, and could be used to help facilitate facial diagnosis in clinical settings. We also investigate creating transitions between averages and exaggerated caricature faces to highlight phenotype differences between patient groups.

Second, we investigate the classification of eight genetic disorders with shallow and deep representations. We compare shape and appearance descriptors based on local and dense descriptors and report significant improvements upon previous work. Furthermore, we made use of transfer learning and part-based models to train convolutional networks for syndrome classification. Our results show that deep learning can be used in the context of classifying genetic disorders, and is superior to shallow descriptors, despite small training datasets.

Neural networks are prone to learning biases present in training datasets and basing their decisions on them. This is particularly relevant for training on small datasets, as is the case in the domain of genetic disorders. We introduce a bias removal algorithm that aims to overcome this challenge. We report three distinct contributions. First, to ensure that a network is blind to a known bias in the dataset, second, to improve classification performance when faced with an extreme bias, and third, to remove multiple spurious variations from the feature representation of a primary classification task.

Lastly, we introduce a novel image augmentation method for learning a deep face embedding, the "Interpolated Clinical Face Phenotype Space", that aims to describe clinically relevant face variation. Our contributions are two-fold: 1) Interpolations between faces that share a class improve deep representation training from small datasets. 2) Between-class interpolations that model the space between classes improve the generalisation performance of the deep representation to unseen syndromes.

Facial Detection of Genetic Disorders



Mohsan Shahzad Alvi

Lady Margaret Hall

University of Oxford

Supervised by Dr Christoffer Nellåker
and Prof Andrew Zisserman

Submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

Trinity 2019

Abstract

An estimated 400,000 children are born every year with rare genetic disorders that significantly affect their quality of life. Early detection and intervention can significantly improve the quality of life of these children. Craniofacial characteristics contain highly useful information for clinical geneticists for diagnosis. This thesis investigates the use of computer vision to aid in the automatic detection of genetic disorders from ordinary facial photographs. This is a non-trivial task, in part due to patient privacy concerns and the scarcity of training data. In the following, we present several approaches to overcome these challenges.

First, we present a method for creating realistic-looking average faces for individuals sharing a syndrome. These averages remove identifiable features, but retain clinically relevant phenotype information and preserve facial asymmetry. This procedure is completely automated, removing the need to expose patient identities at any point during the process, and could be used to help facilitate facial diagnosis in clinical settings. We also investigate creating transitions between averages and exaggerated caricature faces to highlight phenotype differences between patient groups.

Second, we investigate the classification of eight genetic disorders with shallow and deep representations. We compare shape and appearance descriptors based on local and dense descriptors and report significant improvements upon previous work. Furthermore, we made use of transfer learning and part-based models to train convolutional networks for syndrome classification. Our results show that deep learning can be used in the context of classifying genetic disorders, and is superior to shallow descriptors, despite small training datasets.

Neural networks are prone to learning biases present in training datasets and basing their decisions on them. This is particularly relevant for training on small datasets, as is the case in the domain of genetic disorders. We introduce a bias removal algorithm that aims to overcome this challenge. We report three distinct contributions. First, to ensure that a network is blind to a known bias in the dataset, second, to improve classification performance when faced with an extreme bias, and third, to remove multiple spurious variations from the feature representation of a primary classification task.

Lastly, we introduce a novel image augmentation method for learning a deep face embedding, the "Interpolated Clinical Face Phenotype Space", that aims to describe clinically relevant face variation. Our contributions are two-fold: 1) Interpolations between faces that share a class improve deep representation training from small datasets. 2) Between-class interpolations that model the space between classes improve the generalisation performance of the deep representation to unseen syndromes.

Statement of Originality

I hereby declare that, except where made explicitly clear, the contents of this dissertation are my own original work, and to the best of my knowledge do not contain materials submitted in whole, or in part, for consideration for any other degree or qualification at the University of Oxford or any other academic or professional institution.

Mohsan Shahzad Alvi
May 5, 2020

Acknowledgements

This thesis is a product of the guidance, support and friendship of many people that I have come to know during my time at Oxford, and to whom I would like to express my gratitude.

I would like to thank my supervisors Dr. Christoffer Nellåker and Prof. Andrew Zisserman for their continuous guidance, patience and kindness throughout these past years. This thesis would not have been possible without Christoffer's enthusiasm, optimism and dedication, or Andrew's endless stream of brilliant ideas.

I would like to thank Mitchell for being a great colleague and collaborator. Special thanks go to Aravindh, who was generous both in sharing his vast knowledge in computer vision and kind words. I am grateful to Carlos, Weidi, Li, Sam, and João for readily providing guidance when called upon. Thank you to the members of VGG, especially Amir, Yujie, Joon, Daffy, Sophia, Ankush, and Elliot for all the fun we had in the past years. I would also like to thank Felix, Marco, and Olli, for the camaraderie and support during the final stretch.

I would like to express my gratitude towards my family, Abbu-ji, Ammi-ji, Papa-ji, Mama-ji and Ami, to whom I am deeply indebted for their continuous love, patience, and support. Special thanks go to my brother, Ahsan, for his companionship from the very first breath. Finally, I'd like to thank my wife, Amna, for her love and constant support, for keeping me sane over the past few months, and for being my best friend. This thesis is as much your hard work as it is mine.

This thesis would not have been possible without the financial support of the Oxford Centre for Doctoral Training in Healthcare Innovation studentship (EP/G036861/1), for which I am grateful.

Contents

1 Introduction **1**

1.1 Motivation 1
1.2 Key Challenges 3
1.3 Thesis Outline and Contributions 5
1.4 Publications 6

2 Background and Related Work **8**

2.1 Facial Detection of Genetic Disorders 9
2.1.1 Early Work with Significant Manual Input 9
2.1.2 3D Photography and Scans 10
2.1.3 Studying Healthy Facial Variation 11
2.1.4 Automated Approaches 13
2.2 Deep Learning with Convolutional Neural Networks 15
2.2.1 Network Architectures 16
2.2.2 Transfer Learning and Domain Adaptation 17
2.2.3 Image Augmentation 18
2.2.4 Learning Deep Embeddings 20
2.3 Computer Vision Applications to Faces 22
2.3.1 Detecting Face Landmarks 22
2.3.2 Face Verification and Classification 23
2.4 Face Datasets 25
2.4.1 Clinical Face Datasets 25
2.4.2 Deciphering Developmental Disorders study 32
2.4.3 Other Datasets 32
2.4.4 VGGFace2 Dataset 33

3 Creating Average Faces **35**

3.1 Background 36
3.2 Averaging Algorithm 37
3.2.1 Face Detection and Annotation 38
3.2.2 Creating the Shape Average 38
3.2.3 Creating the Appearance Average 42

3.3 De-identification . . . 45
3.3.1 Quantifying De-identification . . . 46
3.4 Results on Clinical Datasets . . . 47
3.5 Creating Transitions Between Faces . . . 51
3.6 Facial Phenotype Caricatures . . . 51
3.7 Conclusions . . . 55

4 Classification of Genetic Disorders 56
4.1 Classifying Syndromes with Engineered-Feature Based Representations 57
4.1.1 Previous Work . . . 58
4.1.2 Improving the Shape Descriptor . . . 60
4.1.3 Improving the Appearance Descriptor . . . 63
4.1.4 Classification Experiments . . . 66
4.1.5 Results and Discussion . . . 67
4.2 Classifying Syndromes using Deep Representations . . . 71
4.2.1 Networks and Pre-training . . . 72
4.2.2 Fine-tuning for Syndrome Classification . . . 73
4.2.3 Part-based Models . . . 74
4.2.4 Image Augmentation . . . 74
4.2.5 Results and Discussion . . . 76
4.3 Conclusions . . . 78

5 Removing Biases from Deep Neural Network Embeddings 80
5.1 Related Work . . . 82
5.2 Datasets . . . 84
5.2.1 LAOFIW: Labeled Ancestral Origin Faces in the Wild . . . 84
5.2.2 "Cleaned" IMDB Dataset . . . 86
5.2.3 Annotated Faces in the Wild . . . 88
5.3 Unbiased Learning Algorithm . . . 88
5.4 Experiments on Removing Bias . . . 91
5.4.1 Removal of a bias from a network . . . 92
5.4.2 Removal of an extreme bias from a network . . . 92
5.4.3 Simultaneous removal of multiple spurious variations . . . 93
5.4.4 Implementation details . . . 93
5.5 Results . . . 94
5.5.1 Removal of a bias from a dataset . . . 94
5.5.2 Removal of an extreme bias from a network . . . 96
5.5.3 Simultaneous removal of multiple spurious variations . . . 97
5.6 Conclusion . . . 101

6 Interpolated Clinical Face Phenotype Space **103**
6.1 Datasets . . . 105
6.2 Interpolating Between Faces . . . 108
6.2.1 Sigmoid-Weighted Appearance Blending . . . 109
6.3 Interpolating Between Faces to Improve Classification . . . 112
6.4 Methods for Learning Face Embeddings . . . 113
6.4.1 Siamese Network with Curriculum Learning . . . 113
6.4.2 Interpolating Between Classes with Triplets . . . 114
6.5 Interpolated Face Embeddings . . . 115
6.5.1 Experiments . . . 115
6.5.2 Implementation Details . . . 116
6.5.3 Results and Discussion . . . 117
6.6 Evaluation on DDD Study Dataset . . . 119
6.7 Conclusion . . . 122

7 Conclusion **124**
7.1 Achievements . . . 124
7.2 Suggestions for Further Work . . . 127

Bibliography **131**

1 Introduction

Contents

1.1 Motivation 1
1.2 Key Challenges 3
1.3 Thesis Outline and Contributions 5
1.4 Publications 6

1.1 Motivation

There are over 7000 known rare diseases, a number that is constantly increasing (Amberger et al., 2015). A rare disease is defined in Europe as one that affects 1 in 20,000 individuals (Smith, 2015). Rare diseases can have many causes, e.g. changes in genes or chromosomes, infections, allergies and environmental causes. Rare diseases that are caused by changes in genes or chromosomes are referred to as genetic disorders and make up for 80% of known rare diseases. These disorders can be verified by identifying these changes, which can help characterise them and aid diagnosis.

Individually, these disorders are rare, but cumulatively up to 8% of the world's population is affected by genetic disorders (Baird et al., 1988). Approximately 2-5% of children are born with developmental disorders that significantly affect their quality of life (Ropers, 2010; Sheridan et al., 2013). This amounts to 400,000

children every year, given the current global demographics. Some example images of public figures with genetic disorders are shown in Figure 1.1.



Figure 1.1: Example images of public figures with genetic disorders. Left to right, top to bottom: American president Abraham Lincoln was thought to have Marfan syndrome; social media celebrity Adalia Rose was diagnosed with Hutchinson-Progeria syndrome; actors Warwick Davis, Peter Dinklage and Paralympian Ellie Simmonds have been diagnosed with achondroplasia; actor Gaten Matarazzo has cleidocranial dysplasia; actors Lauren Potter and Pablo Pineda are diagnosed with Down syndrome.

Early detection and intervention can significantly improve the quality of life of subjects with disorders (Erhan and Belotserkovsky, 2014). However, only a small minority of patients with suspected developmental disorders receive a clinical diagnosis (Hart and Hart, 2009). This is, in part, due to the cause of many of these disorders remaining unknown (Boycott et al., 2017). Everyone carries a large number of rare variants, not all of which will be pathogenic, which makes isolating the disease-causing variant difficult (MacArthur and Tyler-Smith, 2010).

The most powerful approach for identifying a gene, specific to a disorder, is by comparing multiple unrelated individuals with a shared disorder (Schuurs-Hoeijmakers et al., 2012). The shared difference in genome, can then be associated with the genetic disorder (Vorstman and Ophoff, 2013). In the case of ultra rare genetic disorders, however, finding such subjects is very difficult due to the small number of patients with a given disorder (Rath et al., 2014).

One way to identify these individuals for treatment or genetic studies, is to make use of visible manifestations of genetic disorders. 30-40% of genetic disorders cause deviations from the norm in facial or cranial structure (Hart and Hart, 2009; Shaffer et al., 2016). However, these facial manifestations are compounded by the diverse, healthy variations in facial features between individuals. This means that most clinicians may not be able to recognise all possible variants. Learning to recognise these manifestations forms part of specialist training undertaken by clinicians to become dysmorphologists.

In this thesis, we develop computer vision algorithms to facilitate the study and diagnosis of genetic disorders. Our algorithms are fully automated and can analyse ordinary facial photographs. This allows their use at scale, allowing us to benefit from wide access to mobile phone cameras and high-speed internet. This also means affected individuals and their families can directly submit their images without the need for clinicians to collect data. Furthermore, our methods can be also be used in settings where patient consent may restrict access to individual images.

The methods in this thesis are also applicable to variety of non-clinical face analysis, e.g. when access to large datasets is limited, or when biases are suspected in the training data or feature representations of networks.

1.2 Key Challenges

This section summarises some key challenges when using computer vision for the facial detection of genetic disorders from ordinary photos.

Confidentiality and patient consent. De-identified patient records have been widely used for clinical studies (Kushida et al., 2012). Faces, however, are by definition identifiable. This severely limits the possibility to use them in clinical education and computer vision research as it requires patients to consent to it. This leads to a need for our algorithms to work completely automatically, for this to be used ubiquitously in healthcare systems. Additionally, this illustrates the need for the creation of de-identified faces that capture syndromic phenotypes.

Lack of training data. Deep learning for face recognition has reached human-level accuracies (Taigman et al., 2014). These advancements are in part due to the availability of large-scale face recognition datasets, e.g. Guo et al. (2016); Cao et al. (2018). However, training data at this scale is not available in the domain of genetic disorders. For example, there are only about 114 children with the premature ageing disorder Hutchinson-Progeria in the world (Sinha et al., 2014). Only a fraction thereof will be photographed in a clinical setting, a smaller fraction thereof will be made available with consent for research. This raises several challenges, such as preventing over-fitting and dealing with imbalanced classes.

Biases in training data. Convolutional neural networks are known to encode and amplify trends that are present in training datasets. These biases can be present in multiple settings: 1) When transfer learning from a neural network that was previously trained on large face datasets, that are usually composed of public figures and celebrities. This may lead to inheriting the sex, age and ancestral origin biases of the US film industry, e.g. women are usually younger than their male counterparts. 2) A set of images pertaining to a syndrome may have things in common, e.g. being collected from a single hospital that has blue-coloured walls, or a preponderance of Western individuals, thus making these biases strong signals for classification. These can affect the quality of networks that are trained on these datasets, as they may ignore clinically relevant features in favour of biases.

Large number of known and unknown syndromes. We only had access to images of about 50 syndromes across multiple clinical datasets. That is significantly less than the over 7000 known existing disorders (Rath et al., 2014). This is compounded by the significant numbers of disorders that have yet to be discovered. Creating face representations that can classify only the few disorders that we have training data for severely limits their utility in clinical practice. We therefore

need to create models that encode clinically relevant variation that can generalise beyond the training dataset.

1.3 Thesis Outline and Contributions

Creating Average Faces. (Chapter 3) This chapter presents our work on creating realistic-looking average faces from patient photographs. We show how these can be used to de-identify the patients that form part of the average, whilst retaining clinically relevant facial phenotypes. We also show our algorithm's ability to retain left-right facial asymmetry by measuring the shape and locations of face parts. Furthermore, we present methods for creating transitions and caricatures between averages to highlight the facial phenotypes. Lastly, average faces for 25 genetic disorders, generated from clinical datasets, are presented.

Classification of Genetic Disorders. (Chapter 4) This chapter presents our experiments using engineered-feature based and deep representations for the classification of genetic disorders. We investigate shape and appearance representations using hand-crafted features and data fusion thereof. Then, different deep neural networks are trained, making use of transfer learning and part-based models. We also contrast our methods to previous work in the field and show that our methods significantly outperform them. Furthermore, we show that deep representations fair better than engineered-feature based representations.

Removing Biases from Deep Neural Network Embeddings. (Chapter 5) In this chapter, we describe an approach to removing biases from the embeddings of deep neural networks. We introduce a new dataset that we curated, Labeled Ancestral Origin Faces in the Wild, that consists of 14,000 images of faces from four different ancestral origins. We demonstrate the implications of dataset bias using a publicly available face dataset to train age and gender classifiers. We demonstrate that our method can improve the generalisation performance of networks trained on biased datasets. We also show that multiple data sources can be used to remove

multiple biases from the feature representation of a deep neural network.

Interpolated Clinical Face Phenotype Space. (Chapter 6) This chapter introduces an image augmentation method that can be used to improve training networks when constrained to using face small datasets. We first demonstrate how how interpolating between faces during training improves the classification accuracies of a deep neural network. We then investigate the creating of an Interpolated Clinical Face Phenotype Space, a face embedding, in which patients sharing disorders lie close together. Then the benefits of interpolating within and between classes to learn a face embedding that can generalise to new, unseen syndromes are investigated. Lastly, we evaluate the learnt embedding on a large-scale clinical study dataset.

1.4 Publications

The work in this thesis resulted in a number of peer-reviewed publication listed below in reverse chronological order:

The following publication resulted from work in Chapter 5.

Mohsan Alvi, Andrew Zisserman, and Christoffer Nellåker. 2018. 'Turning a Blind Eye: Explicit Removal of Biases and Variation from Deep Neural Network Embeddings'. *Proceedings of the European Conference on Computer Vision (ECCV)*..

The following publications resulted from work in Chapter 3.

Margot R. F. Reijnders, Robert Janowski, **Mohsan Alvi**, et al. 2018. 'PURA Syndrome: Clinical Delineation and Genotype-Phenotype Study in 32 Individuals with Review of Published Literature'. *Journal of Medical Genetics* 55(2):104–13.

Margot R. F. Reijnders, Robert Janowski, **Mohsan Alvi**, et al. 2018. 'De Novo

and Inherited Loss-of-Function Variants in TLK2: Clinical and Genotype-Phenotype Evaluation of a Distinct Neurodevelopmental Disorder'. *The American Journal of Human Genetics* 102(6):1195–1203.

Hemant Bengani*, Mark Handley*, **Mohsan Alvi***, et al. 2017. 'Clinical and Molecular Consequences of Disease-Associated de Novo Mutations in SATB2'. *Genetics in Medicine* 19(8):900–908.
(*) denotes authors contributed equally.

Jeremy F. McRae, Stephen Clayton, Tomas W. Fitzgerald, Joanna Kaplanis, Elena Prigmore, Diana Rajan, Alejandro Sifrim, Stuart Aitken, Nadia Akawi, **Mohsan Alvi**, et al. 2017. 'Prevalence and Architecture of de Novo Mutations in Developmental Disorders'. *Nature* 542(7642):433–38.

2

Background and Related Work

Contents

2.1 Facial Detection of Genetic Disorders 9
2.1.1 Early Work with Significant Manual Input 9
2.1.2 3D Photography and Scans 10
2.1.3 Studying Healthy Facial Variation 11
2.1.4 Automated Approaches 13
2.2 Deep Learning with Convolutional Neural Networks . 15
2.2.1 Network Architectures 16
2.2.2 Transfer Learning and Domain Adaptation 17
2.2.3 Image Augmentation 18
2.2.4 Learning Deep Embeddings 20
2.3 Computer Vision Applications to Faces 22
2.3.1 Detecting Face Landmarks 22
2.3.2 Face Verification and Classification 23
2.4 Face Datasets 25
2.4.1 Clinical Face Datasets 25
2.4.2 Deciphering Developmental Disorders study 32
2.4.3 Other Datasets 32
2.4.4 VGGFace2 Dataset 33

This chapter provides an overview of the topics that are discussed in this thesis. First, Section 2.1 provides background about other works to use computer vision to detect genetic disorders. Section 2.2 provides background into deep learning with convolutional neural networks. We review computer vision applications in

the face domain in Section 2.3. Lastly, we present face datasets that were made use of in this thesis (Section 2.4).

2.1 Facial Detection of Genetic Disorders

In this section, we discuss applications of computer vision to the facial detection of genetic disorders. First, we present an overview of early methods that involved significant manual annotations (Section 2.1.1). Thereafter, 3D photography and scanning methods are presented (Section 2.1.2). Lastly, we provide an overview of recent work making use of the advancements in automated facial landmark annotation and deep learning (Section 2.1.4).

2.1.1 Early Work with Significant Manual Input

Facial detection of genetic disorders using computer vision has been studied for over 30 years. These approaches relied on significant manual input during data capture and analysis.

One of the earliest uses of a neural network for detecting genetic disorders made use of fully-connected multi-layer perceptrons (Herpers et al., 1993). Images were converted to grayscale and were manually rotated, scaled and cropped to 55×72 pixels. The flattened image (3,960 pixels) was passed through a hidden layer of 2-10 neurons and an output layer was used to perform the binary classification between 10 images of dysmorphic faces and 21 images of healthy controls with 95% accuracy.

Manual annotation of landmarks on faces have been used to measure the geometric properties of the face and to extract local textural information, the number of which ranged from 32 (Loos et al., 2003) to 48 (Vollmar et al., 2008) landmarks.

Gabor wavelets (Tai Sing Lee, 1996) of different scales and orientations were used to describe grayscale images patches around these landmarks (Loos et al., 2003). Vollmar et al. (2008) investigated the joint use of wavelets around landmarks along with their standardised coordinates to also capture geometric information. This analysis was done jointly for frontal and profile face photographs. Balliu et al. (2014) used pairwise distances between landmarks and the angles between vertices

of the Delaunay triangulation of 48 manually placed landmarks to describe the geometry of syndromic faces. In each case, the datasets that were analysed were relatively small and collected under controlled lighting and pose conditions with uniform backgrounds: 55 images of 5 syndromes (Loos et al., 2003), 200 frontal and profile images of 14 syndromes (Vollmar et al., 2008), and 205 images of 14 syndromes (Balliu et al., 2014). In each case, one image per patient was used.

2.1.2 3D Photography and Scans

The use of three-dimensional (3D) imagery has also been investigated for the purposes of studying genetic dysmorphism. These images are collected using photogrammetric face scanners that record a depth point-cloud and 2D photographs of a subject. These scanners are usually large need to be operated by a professional.

Hammond et al. (2005) used photogrammetric face scanners that captured up to 20,000 automatically measured 3D points on the faces of 317 healthy controls, and 379 individuals with one of four syndromes. These were manually annotated with 21 3D landmarks to guide the creation of a dense surface model (DSM). Pairwise discriminations between these DSMs for syndromes and control images resulted in classification accuracies of up to 98%.

In a following study, Hammond et al. (2012) captured images with 25,000 automatically measured 3D points on the face for 55 individuals with disorders. These were manually annotated with 24 3D landmaks to create DSMs, as in (Hammond et al., 2005). Averages were also created for syndrome subsets and compared to age and sex matched control averages to create face signature heat maps, showing the contraction and expansion relative to the control average.

More recently, Kung et al. (2015) 200,000 point cloud, 12 manually annotated landmarks for alignment. These were then described by registration against a predefined facial template to obtain 10,000 automatically placed quasi-landmarks. Kung et al. (2015) used the quasi-landmarks to create a face space using principal component analysis (PCA).

Lewyllie et al. (2017) used similar methods to (Kung et al., 2015) to automatically place over 7,000 quasi-landmarks on face. They then used partial least squares regression to investigate the effect of 22q11.2 Deletion syndrome on facial morphology. Lewyllie et al. (2017) also created 3D average faces that show how the phenotype diverges from the healthy control average (Figure 2.1).

Three-dimensional imaging is becoming more affordable and portable thanks to improved smartphone cameras. These images can overcome many of the challenges that we face when using two-dimensional images, e.g. variations in pose and lighting. As a result, we anticipate a greater focus on three-dimensional imaging for clinical face analysis in the future.

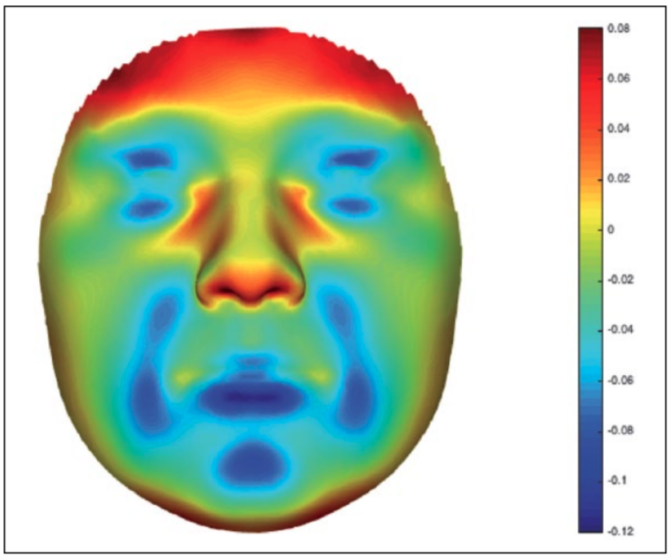


Figure 2.1: Relative inward/outward displacement of facial features of subjects with 22q11.2 Deletion Syndrome, when compared to the average 3D face shape of controls, blue = inward displacement, red = outward displacement (Figure from Lewyllie et al. (2017)).

2.1.3 Studying Healthy Facial Variation

The following concerns the study of healthy variation of facial shape. This is particularly important for disentangling healthy facial variation from that caused by genetic disorders.

Djordjevic et al. (2016) made use of 3D facial images of female twins to study the relative genetic and environmental contributions to facial morphological variation by comparing the facial variation between monozygotic (identical), dizygotic (non-identical) and unpaired twins. The images were taken under restricted head pose, facial expression and lighting conditions. Faces were compared by first manually placing 37 landmarks and comparing principal components and pairwise distances between landmarks. They concluded that genetic factors can explain more than 70% of the phenotypic facial variation in facial size, nose shape, lips prominence and inter-ocular distance.

Genome-wide association studies (GWAS) and 3D facial images find correspondences between facial shape and genetic information (Claes et al., 2018). The authors aimed to study the genetic basis of human facial shape in order to predict facial shape from genome data. Hierarchical spectral clustering (Ng et al., 2002) was used on 10,000 3D measure points on the face to partition the face into global-to-local segments, shown in Figure 2.2. Common shape differences in each segment were corresponded back to genome data.

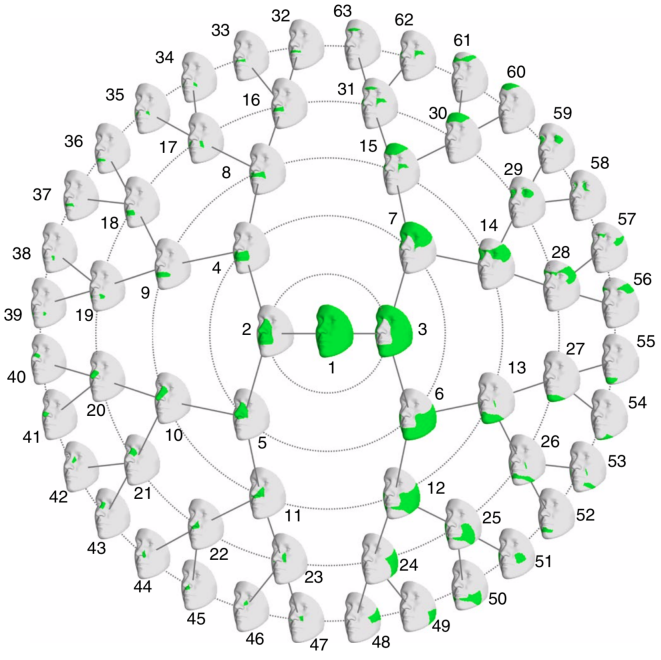


Figure 2.2: Global-to-local facial segments found using hierarchical spectral clustering (Figure from Claes et al. (2018)).

2.1.4 Automated Approaches

The manual nature of the methods in sections 2.1.1 and 2.1.2 limits their practical application. This is due to the requirement for images to be taken under the same lighting and pose conditions with selected cameras, and the manual annotation effort that must be undertaken by an expert. High-quality 2D digital cameras are now affordable and widely available. Improvements in computer hardware also make computer vision algorithms more easy to train and execute. This has lead to the study of automated methods for facial analysis from ordinary photos, that do not require manual annotation.

The Constrained Local Model (CLM) (Cristinacce and Cootes, 2008) and Active Appearance Model (AAM) (Cootes et al., 1998) have been used to automatically annotate facial landmarks on the face. Both of these methods make use of local landmark classifiers and a global template model to constrain the relative positions of landmarks.

These were used to automatically annotate 44 landmarks on the face and to extract local geometric and texture features (Zhao et al., 2014; Kruszka et al., 2017a,b, 2018). These were used to conduct classification between individuals with disorders and gender and age matched controls using support vector machines (SVM).

Local Binary Patterns (Ahonen et al., 2004) were computed around 130 automatically detected facial landmarks and used as features to recognise Cornelia de Lange syndrome (CdLS) (Basel-Vanagaite et al., 2016). A wavelet-based face detector (Viola and Jones, 2004) was used to initialise a chains model (Karlinsky et al., 2010) to detect the locations of the 130 facial landmarks. Distances between landmarks and the extracted local binary patterns around them were used to compare CdLS and healthy controls. The algorithm significantly outperformed the detection rate of experts.

Data fusion between shape and appearance description was also investigated by Ferry et al. (2014). Faces were described by concatenating all pairwise distances between 36 automatically placed landmarks (the shape descriptor) and grayscale values around 13 facial landmarks (the appearance descriptor). These were jointly

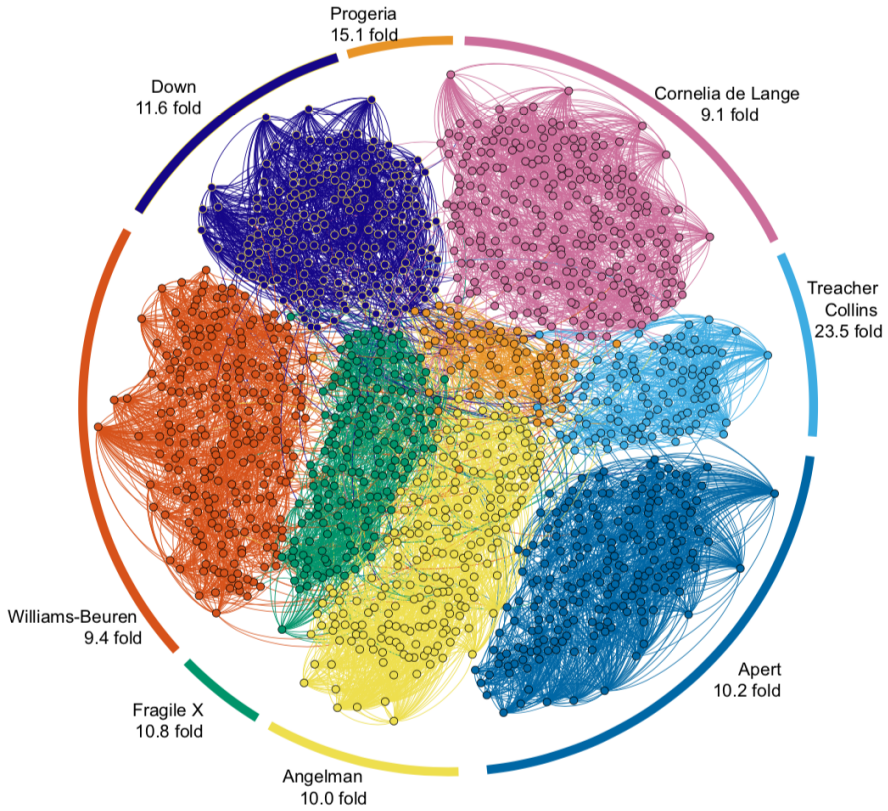


Figure 2.3: Two dimensional representation of the Clinical Face Phenotype Space, showing the separation of different dysmorphic syndromes. The CIF-factor, that is, the X-fold better clustering than random expectation, is also given with each syndrome. (Figure from Ferry et al. (2014))

used to train a discriminant embedding, the 'Clinical Face Phenotype Space', that models human facial dysmorphisms (Figure 2.3). A key contribution of their work was the investigation into generalising to unseen syndromes beyond the training dataset. This work made use of 1,363 facial images across 8 syndromes and 1,515 healthy controls.

Advancements in transfer learning of deep neural networks has led to researchers making use of deep feature representations for syndrome classification. Shukla et al. (2017) investigated the use of deep representations trained for unrelated tasks and appropriated them for the task of classifying genetic disorders. They made use of an AlexNet trained on ImageNet to extract features and classified these

using SVMs. Deep representations were extracted for different parts of the face and concatenated, before classification. 1,126 images of 6 syndromes were collected and classified with 98.8% accuracy with this method.

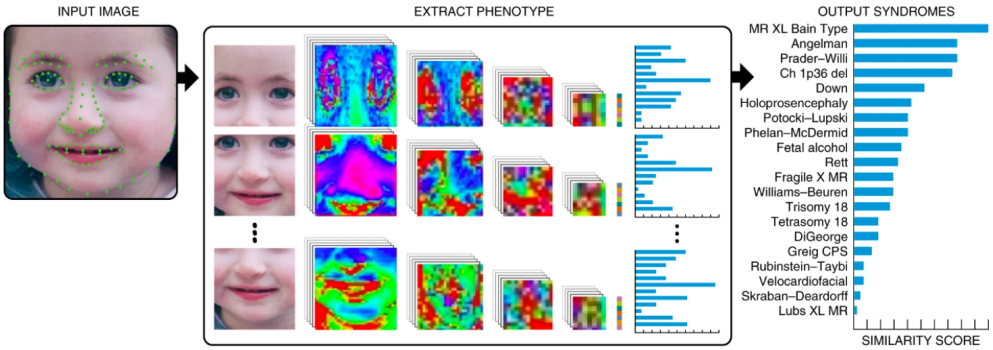


Figure 2.4: DeepGestalt part-based deep learning approach: Networks are trained to classify each face part and jointly vote for the final classification. (Figure from (Gurovich et al., 2019))

Transfer learning was also used by Gurovich et al. (2018, 2019) to train their algorithm, named DeepGestalt, is also a part-based model, like (Shukla et al., 2017). A large face recognition dataset (Yi et al., 2014) was used to pre-train different networks for each face part and then fine-tuned on 26,190 syndromic images. At inference, the outputs from individual face-part networks were used to jointly predict 215 syndromes from a test set of 502 images (Figure 2.4). This work was run by FDNA[1], a company that is also looking to answer many of the same research questions that are presented in this thesis.

2.2 Deep Learning with Convolutional Neural Networks

Convolutional neural networks (CNN) are the state-of-the-art for image classification and have outperformed engineered-feature based methods on public benchmarks (Chatfield et al., 2014). They are a type of deep neural network, that contain layers with convolutional filters, instead of only assigning unique per-pixel weights as in

[1]https://www.fdna.com/

fully-connected networks. A convolutional filter is defined by a set of trainable weights that slides across the whole input, at each point convolving with a local image patch. The output is then fed forward to the next layer of the network. These filters are trained to extract similar patterns, or features, in data that are spatially connected, e.g. different parts of the image. A typical CNN consists of a number of convolutional layers, followed by several fully-connected layers. These layers extract low-level features that are translated into higher level concepts layer-by-layer. The fully-connected layers can be thought of as conducting inference on features extracted by the convolutional layers.

An introduction to different CNN architectures is given in Section 2.2.1. The next two sections pertain to methods that are particularly useful for training CNNs with smaller training datasets (Sections 2.2.2 and 2.2.3). Lastly, a summary of methods for learning deep embeddings are detailed in Section 2.2.4.

In the context of this thesis, deep learning will refer to the use of CNNs.

2.2.1 Network Architectures

The AlexNet architecture (Krizhevsky et al., 2012) was one of the first widely used CNNs for image classification and achieved state-of-the-art performance on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) (Deng et al., 2009). It consists of five convolutional layers and two fully-connected layers.

Simonyan and Zisserman (2014) showed that CNNs could be improved by increasing their depth. Their VGGNet architecture used small 3×3 convolutional filters and 19 layers and significantly improved upon the AlexNet architecture when evaluated on ILSVRC. The VGG-M architecture of Chatfield et al. (2014) aims to strike a balance between greater performance and faster training times and consists of five convolutional layers and three fully connected layers.

Network depth has been directly correlated with improving accuracy of CNNs (He and Sun, 2014; Simonyan and Zisserman, 2014), however, researchers found that they are more challenging to train (He and Sun, 2014; Srivastava et al., 2015). He et al. (2015) investigated networks with up to 152-layer networks, $8\times$ deeper than

VGG nets. They overcome the difficulties in training this model by introducing skip connections around residual blocks (Figure 2.5). Their ResNet50 variant is a 50-layer residual network that has been widely adopted for image classification.

Squeeze-and-Excitation networks (SENet) (Hu et al., 2017) are an extension of the ResNet architecture where inter-dependencies between channels are modelled. ResNet and SENet are currently considered the best classification architectures.

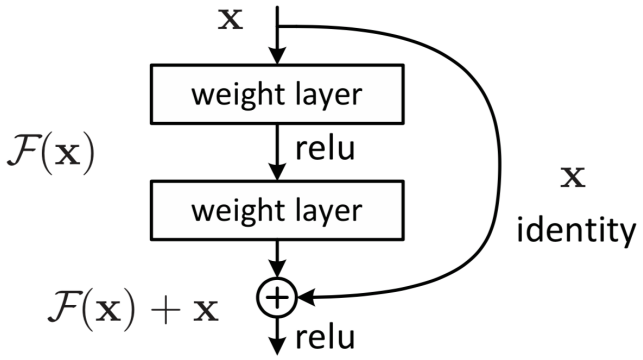


Figure 2.5: Residual block of ResNet architecture. Figure from (He et al., 2015).

2.2.2 Transfer Learning and Domain Adaptation

Deep learning relies on large datasets for training, which are not always available. Transfer learning aims to circumvent this limitation where domain-specific training datasets would usually be considered too small (Shao et al., 2015; Girshick et al., 2013; Long et al., 2015; Oquab et al., 2014). Networks that are trained on large datasets in other domains, such as ImageNet (Deng et al., 2009) and VGGFace2 (Cao et al., 2018) can be trained further to build a domain-specific classifier. This does, however, limit us to using the pre-existing architecture. Pre-training refers to an architecture of choice being trained on a large dataset to initialise the domain-specific training task.

Domain adaptation is a field of research that aims to adapt a feature representation for a task in a source domain to the same task in a target domain, e.g. if trained on images in daylight, generalising to images at night-time (Ganin et al., 2017; Tzeng et al., 2015; Tzeng et al., 2017). An illustration of this approach is

given in Figure 2.6. One approach is to use domain discriminators to measure whether a network can identify the source domain of a given image (Tzeng et al., 2015; Tzeng et al., 2017). The rationale behind this method, is that making the feature representation domain agnostic will make it more transferable. A benefit of this approach is that the target domain need not be exhaustively labelled.

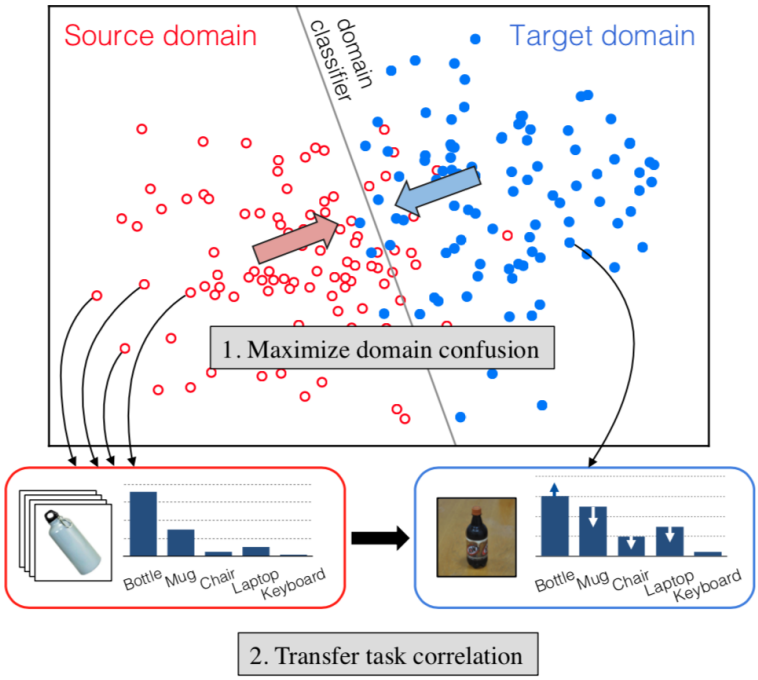


Figure 2.6: Domain adaptation by introducing domain confusion loss during training. (Figure from (Tzeng et al., 2015)).

2.2.3 Image Augmentation

The parameters that define these layers are learned directly from the training data. A typical CNN contains a large number of parameters, e.g. the VGG-M framework contains over 100 million (Chatfield et al., 2014), but large datasets often only contain around 1 million images (Deng et al., 2009). One of the main challenges during training with such a large number of tunable parameters is to avoid over-fitting. Image augmentation is one approach for overcoming this challenge.

Image augmentation aims to artificially increase the size and diversity of the training dataset. There are two forms of augmentation. One that creates a new



Figure 2.7: Random combinations of image augmentations that are used when training CNNs. These examples contain random crops, rotation, horizontal flips, grayscale, blur, noise, changes in exposure, and JPEG compression.

instance of a single class. This can be done by applying geometric and colour transformations on an image, by adding noise or blurring, and partial occlusion. Second, images from different classes can be mixed by manipulating pixel values or by feature space manipulation to generate mixtures with generative models. Figure 2.7 demonstrates a random combination of typical image augmentations that would be applied to an image at training time.

In the following, we present some popular image augmentation techniques.

Geometric transformations. These can include rotations, random crops, horizontal and vertical flipping, affine transformations, warping, and translations. Each of these return a slightly changed perspective of an image. Not all geometric augmentations are always appropriate, e.g. training a face classifier with upside-down face augmentations may deteriorate classifier performance. It may be more appropriate to ensure faces to be upright during pre-processing. Anisotropic transformations change the shape of a face, which is a strong signal for detection genetic disorders and are therefore avoided.

Color transformations. The appearance of an images can be adjusted to reflect

different lighting conditions, by increasing and reducing its exposure and contrast. Additionally, noise can be added to colour channels to simulate differences in white-balancing. Even though colour images contain a richer signal for classification, converting images to grayscale allows networks to generalise to test cases of images without colour.

Noise, Blurring, Compression. The human eye is far better at identifying objects in noisy images than computer vision models, something that is confirmed by the wide-spread use of CAPTCHA graphics. Gaussian or motion blurs and different types of noise can be added to images to make the network's feature representation more robust. The same is true for image compression. JPEG compression is used at point of capture and in post-processing to reduce the file-size of images. This does not significantly affect their appearance to the human eye, but it does change the appearance to the neural network.

Mixup. The image augmentations mentioned so far aim to model the within-class distribution given different appearances of a training image. The Mixup algorithm (Zhang et al., 2017) aims to extend the training data to samples between classes. It assumes that linear interpolations in the feature representation of images should map onto linear interpolations of the associated targets. This method has led to improved classification performance and was shown to improve the robustness of representations (Zhang et al., 2017).

2.2.4 Learning Deep Embeddings

An image embedding is a low-dimensional representation of an input image. In the supervised case, we assume access to a training dataset that contains images and their class labels. The objective would be to learn an embedding that captures information about the relationship between the images and their classes, such that images of a shared class are placed close together in the embedding space.

An embedding, in the context of a neural network, is a feature representation of an image that has been trained for a specific task. This is usually the output of a fully-connected bottleneck layer for a classification network, of the output of the final layer for an embedding network. Three methods for creating embeddings are detailed below: 1) siamese networks, 2) prototypical networks, and 3) triplet networks.

Learning an Embedding with a Siamese Network

A siamese network (Chopra et al., 2005) compares the embeddings of pairs of images (x_1, x_2) using a contrastive loss function (Equation 2.1). At each instance, two images are passed through the network and their distance in embedding space, $D(x_1, x_2)$, is compared. The objective is to create an embedding space, in which positive pairs ($y = 1$) of images of the same class lie closer to each other than negative pairs ($y = 0$) of images for different classes by at least a margin m.

$$L(x_1, x_2, y) = yD(x_1, x_2) + (1 - y)\max(0, m - D(x_1, x_2)) \tag{2.1}$$

Learning an Embedding with a Prototypical Network

Prototypical networks (Snell et al., 2017) are used in instances where only small amounts of training data are available. They aim to learn an embedding in which all members of a class cluster around a prototype representation of that class. At each iteration, K classes are selected. A prototype c_k for class k is calculated from a subset of images belonging to that class, S_k , called the support examples (Equation 2.2). Predictions are then made for a set of query images q, by finding the closest prototype. The loss function is the softmax loss over the distance between the correct prototype and a query image (Equation 2.3).

$$c_k = \frac{1}{|S_k|} \sum_{i=1}^{|S_k|} f(x_i), \tag{2.2}$$

where $f(x_i)$ is the embedding of image x_i.

$$L(x, y = k) = -\log\left(\frac{e^{D(f(x),c_k)}}{\sum_{i=1}^{K} e^{D(f(x),c_i)}}\right) \tag{2.3}$$

Learning an Embedding with a Triplet Network

Triplet learning was introduced by (Schultz and Joachims, 2004) where relative comparisons such as "A is closer to B than A is to C" are used to learn a distance metric. More recently, it has been used successfully in deep learning for face verification (Schroff et al., 2015).

Triplet networks are trained by sampling three images (x, x_+, x_-), where the anchor image x and positive image x_+ are from the same class and the negative image x_- is from a different class. The triplet loss minimizes the distance in embedding space between the anchor and the positive, and maximizes the distance between the anchor and the negative (Figure 2.8).

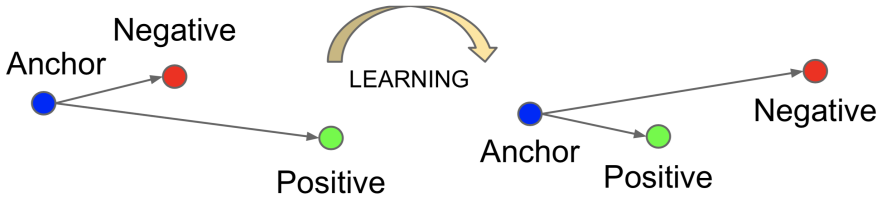


Figure 2.8: Triplet sampling method for training face embeddings (Figure from (Schroff et al., 2015))

2.3 Computer Vision Applications to Faces

This section provides an overview of computer vision applications in the face domain.

2.3.1 Detecting Face Landmarks

The field of face annotation aims to localise a set of feature points on a facial photograph. These are usually chosen at key locations on the face, e.g. the eyes, nose, mouth, and jaw-line. The training dataset is usually pre-annotated with ground truth locations for landmarks.

Active Appearance Models (AAM) (Cootes et al., 1998; Edwards et al., 1998) combine SVM classifiers for each landmark with a statistical model of shape variation to decide on the best constellation of landmarks.

Iterative methods initialise a shape constellation and update the estimated landmark positions at each iteration based on local features. The Supervised Descent Method (SDM) does so by treating the annotation task as a minimisation problem, where the difference between feature representations of predicted landmarks and their true locations, described using the SIFT descriptor (Lowe, 2004), must be minimised (Xiong and De la Torre, 2014). They do so by learning the parameters for an approximation of Newton's method directly from the dataset. Another iterative approach was taken by Kazemi and Sullivan (2014) to learn an ensemble of regression trees. Starting from an initial average shape, at each iteration, an update vector is predicted that is added to the current shape. The landmarks are predicted based on image features and assigned an index based on the shape estimate at that iteration. This allows for an element of shape invariance, which as the algorithm progresses, reduces when the locations of landmarks become more certain.

Deep methods have also been used in cascades, where the output of earlier networks is a more coarse prediction and following networks make finer adjustments. A three-level cascade model was investigated by Sun et al. (2013). They also use local weight sharing to learn more effective regional properties, e.g. eyes are semantically different to the mouth.

2.3.2 Face Verification and Classification

Face verification and classification is based on learning a representation that is invariant to expression, age, pose, lighting, and occlusions. This representation can then be compared to other representations. To verify that two images contain the same individual, their feature representations must lie close to each other in the face representation space (Figure 2.9).

Face descriptors using Fisher vector encodings were investigated by Simonyan et al. (2013). A Gaussian mixture model was learned to encode dense SIFT descriptors of faces in the training data. This descriptor was further improved by discriminant metric learning. This method was the state-of-the-art for face recognition using engineered-feature based representations.

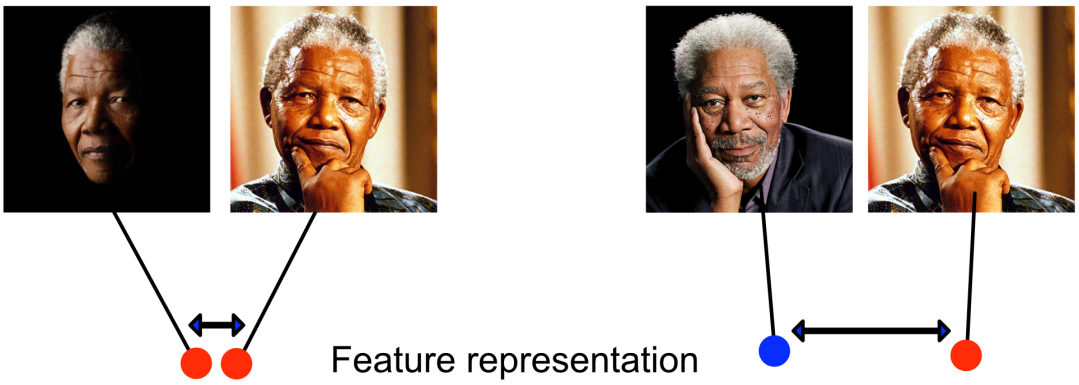


Figure 2.9: Face verification example: Feature representations are extracted for two pairs of images of two identities (red, blue). If the features lie within a distance threshold, they are considered to be of the same identity.

Recent advancements in convolutional neural networks and access to large-scale training datasets have led to a number of deep approaches to face verification. DeepFace uses a frontalization algorithm to normalise the inputs to a deep network consisting of two convolutional layers, three locally-connected layers and two fully-connected layers (Taigman et al., 2014). The frontalization algorithm is guided by a landmark annotation algorithm followed by 2D alignment to a generic shape model. They also make use of the representation two compare individuals using a learned, weighted distance metric using linear SVMs. The network and a frontalization example are shown in Figure 2.10

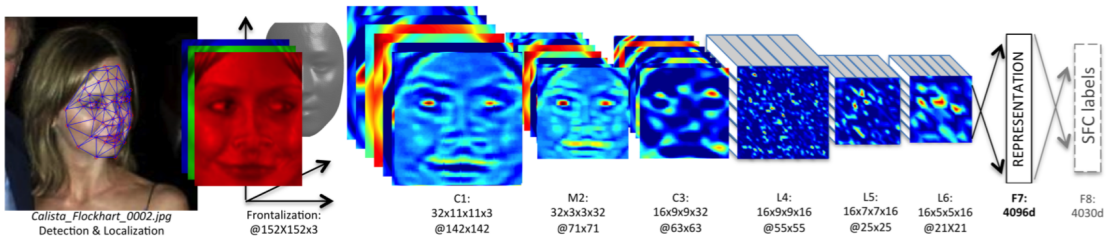


Figure 2.10: Outline of DeepFace architecture. The face is first detected and annotated with landmarks that are used to frontalize the face with the use of a 3D model. The network consists of two convolutional layers, three locally-connected layers and two fully-connected layers. The penultimate layer represents the deep feature representation of the face (Figure from Taigman et al. (2014))

Facenet is trained to directly optimize a compact 128-dimensional embedding,

instead of an intermediate bottleneck layer (Schroff et al., 2015) using an Inception network architecture (Szegedy et al., 2014). The network was trained using triplets (Section 2.2.4). The triplet network learns an embedding in which individuals of shared identity are closer to each other in embedding space that those of different identities. The network was trained on 100M-200M images of 8M individuals and achieved state-of-the-art performance on the LFW dataset (Huang et al., 2007). Parkhi et al. (2015) trained a VGG-16 network (Simonyan and Zisserman, 2014) on the VGGFace dataset (Section 2.4.3), which, even though it contained significantly fewer images (2.6M for 2.6k identities), also achieved comparable results to Schroff et al. (2015). This work was further extended by Cao et al. (2018), who trained Resnet50 (He et al., 2015) and SENet50 (Hu et al., 2017) networks, on the VGGFace2 dataset (Section 2.4.4).

2.4 Face Datasets

We now present the face datasets that were used for the results in this thesis. We will describe the clinical face datasets we used for syndrome classification followed by two large-scale face recognition datasets, VGGFace and VGGFace2, that were also used to pre-train networks, which were used to train the learning algorithms.

2.4.1 Clinical Face Datasets

Publicly available datasets containing faces of individuals with genetic disorders are scarce. This is in part due to the nature of these disorders being rare, and in part due to confidentiality restrictions.

We used three clinical face datasets to train and evaluate algorithms for the facial detection of genetic disorders, which we will discuss in turn: the previously published LSITW dataset, the Minerva Image Resource, the Patients in PubMed Publication dataset, and the Deciphering Developmental Disorders Dataset.

There are a small number of public figures that are known to have genetic disorders, some of which are shown in Figure 1.1. These images are representative of

the types of images in the clinical datasets that we used for training the algorithms in this thesis, which we cannot publish due to patient privacy concerns.

LSITW: Labelled Syndromes in the Wild

The Labelled Syndromes in the Wild (LSITW) dataset was procured by Ferry et al. (2014) for the purpose of classifying genetic disorders from facial photographs. The dataset consists of 1,363 images of eight genetic disorders and 1,515 healthy controls. The images were collected from publicly available sources on the internet. The dataset is diverse in terms of sex, age, ancestral origin, facial expression, lighting, pose and image background.

A detailed summary of the dataset is given in table 2.1. The training data consists of 80% of the images for each class and the remaining 20% of images are held out for testing.

Table 2.1: List of images and syndromes in LSITW dataset

Syndrome	Images
Angleman	205
Apert	203
Cornelia de Lange	179
Down	199
Fragile X	164
Progeria	78
Treacher Collins	103
Williams-Beuren	232
Healthy Controls	1,515

Minerva Image Resource

The Minerva Image Resource (MIR) is a soon to be released dataset collected and maintained by the Minerva Initiative (Nellåker et al., 2019). The Minerva Initiative aims to enable more research in the field of diagnosis by means of image analysis. The MIR is a repository of images from a number of clinical collaborations. Apart

from facial photographs, images of other body parts are also present. We used a face detector to filter out the images that did not contain faces, resulting in 1,367 images.

Some of the syndrome groups in this dataset consist of only a few individuals, and were thus too small to be used in our experiments. We only made use of syndromes that were shared by more than 10 patients. Table 2.2 summarises the remaining 1,244 images and syndromes that were used in this thesis.

Disorder / Gene	Images
16p11.2 Del	55
16p11.2 Dup	33
17q21.31 microdeletion syndrome	149
Achondroplasia	10
B-Raf proto-oncogene, serine/threonine kinase	39
Coffin-Lowry syndrome	22
Cornelia de Lange syndrome	270
Costello syndrome	11
Down syndrome	38
Gorlin syndrome	26
Hutchinson-Gilford progeria syndrome	10
KRAS proto-oncogene, GTPase	11
Kleefstra syndrome due to 9q34 microdeletion	133
Lenz-Majewski hyperostotic dwarfism	18
Monosomy 5p	15
Myeloid neoplasms associated with eosinophilia and abnormality of PDGFRA, PDGFRB or FGFR1	10
Noonan syndrome	111
Orofaciodigital syndrome	10
Otopalatodigital syndrome	16
PGAP3	31
PIGA	39
PIGN	14
PIGT	21
PIGV	36
PTPN11	19

Rothmund-Thomson syndrome	11
SOS Ras/Rac guanine nucleotide exchange factor 1	32
Sotos syndrome	15
Waardenburg syndrome	19
protein tyrosine phosphatase, non-receptor type 11	20

Table 2.2: List of images and syndromes in Minerva Image Resource

Patients in PubMed Publications

Patients in PubMed Publications (PiPP) is a dataset that has been curated by means of mining medical publications of rare disorders (Dawson et al., 2017). PubMed is an online resource with over 30 million citations for biomedical literature. Some of these publications contain facial photographs of individuals with rare disorders. Dawson et al. (2017) developed an algorithm that can harvest facial photographs from publications pertaining to a syndrome. The images were then manually verified to be examples of the queried disorders. The image quality varies significantly across the dataset. Some images from older publications, that were presumably digitised by scanning, are grainy and monochrome, whereas newer publications contain higher quality photographs. We only used disorders that contained more than 10 patients, resulting in 1501 images which are detailed in table 2.3.

Disorder / Gene	Images
17p13.3 microduplication syndrome	15
17q21.31 microdeletion syndrome	17
7q11.23 microduplication syndrome	36
Aarskog-Scott syndrome	39
Acrofacial dysostosis	17
Angelman syndrome	151
Apert syndrome	20
Auriculocondylar syndrome	12

Cardiofaciocutaneous syndrome	50
Cleidocranial dysplasia	31
Coffin-Lowry syndrome	34
Craniofacial-deafness-hand syndrome	22
Craniosynostosis	11
DYRK1A-related intellectual disability syndrome	15
Ehlers-Danlos syndrome	71
Ehlers-Danlos syndrome, dermatosparaxis type	19
Ehlers-Danlos syndrome, musculocontractural type	25
Fragile X syndrome	86
Frontometaphyseal dysplasia	29
Goldenhar syndrome	26
Gorlin syndrome	19
Hemifacial microsomia	12
Hypertelorism, Teebi type	12
Jackson-Weiss syndrome	12
Loeys-Dietz syndrome	25
Mandibulofacial dysostosis	70
Mandibulofacial dysostosis-microcephaly syndrome	19
Marfan syndrome	45
Melnick-Needles syndrome	17
Monosomy 9q22.3	15
Muenke syndrome	25
Nager syndrome	15
Oculodentodigital dysplasia	18
Orofaciodigital syndrome	11
Orofaciodigital syndrome type 1	11
Pitt-Hopkins syndrome	53
Prader-Willi syndrome	93
Proteus syndrome	40
Rett syndrome	13
Saethre-Chotzen syndrome	27
Shprintzen-Goldberg syndrome	38

Treacher-Collins syndrome	80
Waardenburg syndrome	21
Waardenburg syndrome type 1	17
Williams syndrome	51
X-linked intellectual disability with marfanoid habitus	16

Table 2.3: List of images and syndromes in PiPP database

Summary of Clinical Face Datasets

Table 2.4 summarises all of the clinical datasets that are listed in this section.

Disorder / Gene	Images
16p11.2 Del	55
16p11.2 Dup	33
17q21.31 microdeletion syndrome	149
Aarskog-Scott syndrome	10
Achondroplasia	10
Acrofacial dysostosis	16
Angelman syndrome	224
Apert	203
Auriculocondylar syndrome	10
B-Raf proto-oncogene serine/threonine kinase	39
Cardiofaciocutaneous syndrome	32
Cleidocranial dysplasia	12
Coffin-Lowry syndrome	40
Cornelia de Lange syndrome	449
Costello syndrome	11
Down syndrome	237
Ehlers-Danlos syndrome	26
Fragile X syndrome	218
Goldenhar syndrome	12
Gorlin syndrome	40

Healthy Controls	1515
Hutchinson-Gilford progeria syndrome	88
KRAS proto-oncogene GTPase	11
Kleefstra syndrome due to 9q34 microdeletion	133
Lenz-Majewski hyperostotic dwarfism	18
Mandibulofacial dysostosis	31
Marfan syndrome	32
Monosomy 5p	15
Muenke syndrome	14
Myeloid neoplasms associated with eosinophilia and abnormality of PDGFRA, PDGFRB or FGFR1	10
Nestor-Guillermo progeria syndrome	10
Noonan syndrome	111
Orofaciodigital syndrome	24
Otopalatodigital syndrome	26
PGAP3	31
PIGA	39
PIGN	14
PIGT	21
PIGV	36
PTPN11	19
Prader-Willi syndrome	41
Proteus syndrome	15
Rothmund-Thomson syndrome	11
SOS Ras/Rac guanine nucleotide exchange factor 1	32
Saethre-Chotzen syndrome	17
Sotos syndrome	15
Stickler syndrome	10
Treacher-Collins syndrome	150
Waardenburg syndrome	49
Williams syndrome	248
protein tyrosine phosphatase non-receptor type 11	20

Table 2.4: List of all images and syndromes that were used in this thesis.

2.4.2 Deciphering Developmental Disorders study

The Deciphering Developmental Disorders dataset (DDD) is the result of a longitudinal study (Deciphering Developmental Disorders Study et al., 2017) that aims to use new genetic technologies to help diagnose developmental disorders. Patients were recruited to the study in 24 regional Genetics Services across the United Kingdom and Republic of Ireland. It contains genetic and photographic information for over 12,000 undiagnosed individuals.

Due to consent restrictions, we were not given access to these images. The algorithms were, therefore, packaged into command-line scripts that were run "blind". A face detector was used to identify which images from the total set of 14,000 contained faces. These were then annotated using methods from Section 3.2.2. This resulted in 12,706 images of 6,781 patients. Non-identifiable, deep representations were extracted from the images and analysed further.

Diagnosis is only available for a minority of patients. The dataset contains a confirmed diagnosis for 135 patients, and information about genetic variants for 1676 patients.

2.4.3 Other Datasets

We made use of two widely-used, non-clinical face verification datasets, VGGFace and VGGFace2.

VGGFace Dataset

VGGFace is a face recognition dataset consisting of 2.6 million facial photographs of 2,622 public figures (Parkhi et al., 2015). The dataset was procured by downloading publicly available images of popular individuals listed on the Internet Movie Data Base (IMDB)[2] using the Google Image Search engine. Many of these images may not contain the identity in question, so these needed to be removed. To this end,

[2] `https://www.imdb.com`

Parkhi et al. (2015) trained linears SVMs on an engineered-feature based feature representation of the top 50 Google search results for each individual. All images were then ranked by classifier score and only the top images were retained. The images were then manually filtered to remove any further noise. Some example images are shown in Figure 2.11.

The training data consists of 89% of images and the remaining 11% were reserved for testing. All identities occur in both sets.

2.4.4 VGGFace2 Dataset

VGGFace2 is a face recognition dataset consisting of 3.31 million facial photographs of 9,131 public figures (Cao et al., 2018). It consists of publicly available images, of individuals in the Freebase knowledge graph [3] for which a sufficiently large number of images were found using the Google Image Search engine. VGGFace2 contains a greater variety of ages and poses for each identity when compared to VGGFace. This was achieved by appending keywords, such as "sideview" and "very young" to an individual's name. As a result, networks trained on VGGFace2 for face verification were shown to be more invariant to age and pose than those trained on VGGFace (Cao et al., 2018). Figure 2.12 contains some example images, demonstrating the variation in age and pose for several identities in the dataset.

The training dataset consists of 8631 identities and the remaining 500 are reserved for testing.

[3] `https://www.freebase.com`

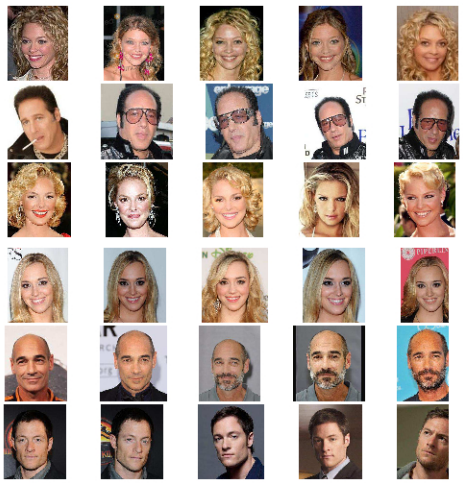


Figure 2.11: Example images from the VGGFace dataset (Parkhi et al., 2015).



(a) Age variation examples **(b)** Pose variation examples

Figure 2.12: Example images from the VGGFace2 dataset (Cao et al., 2018). Two rows of images are shown for each identity to demonstrate (a) age and (b) pose variation in the dataset.

3
Creating Average Faces

Contents

3.1 Background 36
3.2 Averaging Algorithm 37
3.2.1 Face Detection and Annotation 38
3.2.2 Creating the Shape Average 38
3.2.3 Creating the Appearance Average 42
3.3 De-identification 45
3.3.1 Quantifying De-identification 46
3.4 Results on Clinical Datasets 47
3.5 Creating Transitions Between Faces 51
3.6 Facial Phenotype Caricatures 51
3.7 Conclusions 55

A significant number of genetic disorders have manifestations in facial or cranial structure (Hart and Hart, 2009; Shaffer et al., 2016) that can help clinicians in diagnosing patients. However, confidentiality concerns that arise when sharing identifiable patient photographs mean that many clinicians have not seen the facial phenotype of many syndromes. We believe that visual libraries of known phenotypes that do not identify patients could present clinicians with a learning and reference tool to aid diagnosis. This leads us to investigate methods for synthesising de-identified composite faces, which capture the clinically relevant phenotype for a given disorder.

In the following, we present our method for creating realistic-looking average faces that preserve phenotype information and left-right facial asymmetry. The algorithm is completely automated, making use of face detection and annotation algorithms, which enable the creation of average faces without exposing the identity of any individual at any point during the process. This chapter makes use of face detectors and annotators from (King, 2009).

The methods in this chapter are a result of an iterative procedure to improve the quality of the average faces. Section 3.2.2 introduces our shape alignment and averaging algorithm that is robust to out-of-plane rotations. This allows the use of images with a wider range of poses. Averaging faces can result in "averaging out" clinically relevant facial asymmetries, which we account for in Section 3.2.2. Section 3.2.3 describes how we warp the pixel values of individual images to create the average face to improve de-identification. We present results of our methods in Section 3.4. The average faces were constructed using images from several clinical datasets (Section 2.4).

We found that creating transitions between different patient groups highlighted the differences between them. In Section 3.5, we describe how to create transitions that warp the shape and appearance of a face smoothly to another face. Furthermore, in Section 3.6 we discuss the creation of caricature averages, that exaggerate the difference between the syndrome characteristic phenotype and healthy controls.

Lastly, we discuss the ability of our algorithm to de-identify the average by removing identifiable features from each constituent image (Section 3.3). We also present an experiment to quantify the level of de-identification for averages with increasing numbers of members.

3.1 Background

Face averaging has been extensively studied in the context of understanding facial beauty (Langlois and Roggman, 1990; Langlois et al., 1994; Rhodes et al., 1999; Rennels et al., 2008). Early research required images to be taken under controlled environments, with pose, position and lighting being strictly controlled. More

recently, facial landmarks have been introduced as a means to allow for fewer restrictions on the input images, e.g., a software tool called InterFace that allows manipulation of faces and the creation of transitions and averages (Kramer et al., 2017). These face manipulation tools generally rely on manual annotation of, in some cases, over 100 facial landmarks (Rhodes et al., 1999).

Tiddeman et al. (2001) use average faces, that they name prototype images, to define transformations between characteristic prototypes of face attributes e.g. age. These can be used to apply these attributes onto new faces, e.g. to make a young face appear older. Face averaging as a means to mitigate the natural variations between photographs of an individual, e.g. pose, lighting, and expression, was used by Robertson et al. (2015) to improve mobile phone face recognition.

Average faces have been hypothesized to look more attractive, as they average out individual asymmetries and look more symmetrical (Langlois and Roggman, 1990; Iyengar et al., 2015). Certain disorders, however, can cause a phenotype that appears on only one side of the face. When averaging faces naively, one can unwittingly "average out" this phenotype.

Some research groups (Baird et al., 1988; Ferry et al., 2014) and companies[1] have previously investigated average faces for the purpose of visualizing facial phenotypes. Most of these methods, however, rely on manual annotations to create accurate images, greatly limiting their scalability. They also do not look realistic, lacking colour or presenting unnatural facial symmetry.

Averaging 3D faces has also been investigated to study facial archetypes (Shaweesh et al., 2006; Seager et al., 2009). In both papers, individuals were photographed using 3D cameras and the resulting face shells were averaged to assess differences in facial morphologies of different ancestral backgrounds.

3.2 Averaging Algorithm

We introduce a fully-automated, three-step algorithm that can create average faces from facial photographs, such as those shown in Figure 3.1. There are three

[1] `https://www.face2gene.com`

main steps in our algorithm:

1. Face detection and annotation of 68 defined facial landmarks on the face.
2. Creation an average face mesh from multiple individuals with a shared genetic disorder, which we name the shape average.
3. Warping of individual faces onto the average face mesh to create the average face, which we name the appearance average.



Figure 3.1: Sample images that can be input into our averaging algorithm

3.2.1 Face Detection and Annotation

The averaging algorithm depends on accurate face detection and landmark annotations, which are used to create the averages.

We make use of a CNN face detector and a 68-landmark annotation algorithm based on work by Kazemi and Sullivan (2014). Both of these are available as part of the DLIB machine learning toolbox (King, 2009). Figure 3.2 shows the positions of the landmarks that are returned when using this annotator. There are 10 landmarks along the eyebrows, 12 landmarks for the eyes, 10 landmarks for the nose, 19 describing the mouth, and 17 points along the jaw-line.

3.2.2 Creating the Shape Average

The shape average is created from the average locations of facial landmarks for all patients with a shared disorder.

The averaging algorithm is initialized with a reference constellation, which is a symmetrical constellation of annotations for a generic face. We created this reference

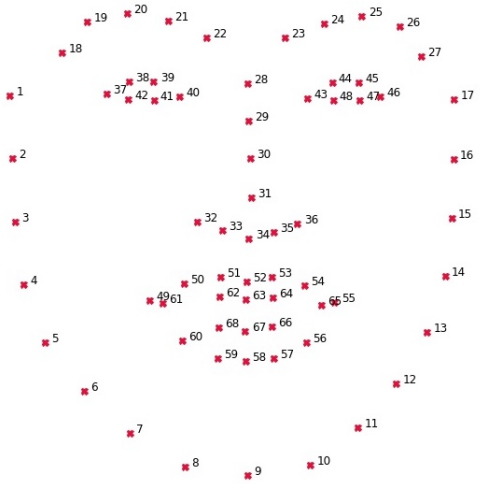


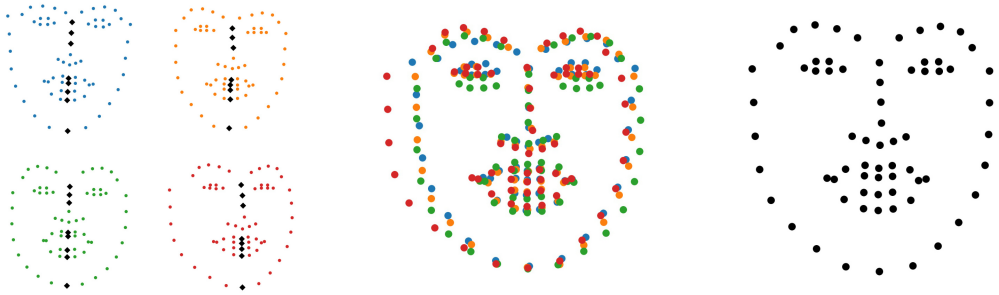
Figure 3.2: The locations of 68 landmarks on the face that were predicted using an annotation algorithm based on Kazemi and Sullivan (2014)

constellation by averaging the landmarks of 2,000 healthy control faces and their horizontal flips at various ages for both sexes. This reference face constellation is used to align faces of varying sizes and orientations to the same coordinate frame via the Procrustes transform. The aligned locations of the annotations for each patient are then averaged to create the shape average. This shape average captures the typical geometry of the face for a given syndrome.

In the following, we present our method for aligning facial landmarks to the reference constellation, followed by how to preserve asymmetrical features.

Robust Face Alignment

The Procrustes transformation is used to translate, rotate and scale landmarks to align maximally with the reference constellation. This transform is sensitive to out-of-plane rotations: if a face is rotated to one side, the locations of landmarks on either side of the face are affected by the foreshortening effect. This is most noticeable along the jaw-line, as can be seen in the bottom-right face in Figure 3.3a. This leads to poor alignment of annotations between faces. To mitigate this, we use only landmarks 28, 29, 30, 31, 52, 58, 63, and 67 (Figure 3.2) along the horizontal



(a) Annotated faces **(b)** Aligned faces **(c)** Shape average

Figure 3.3: (a) Example annotations for four input faces. Landmarks that are used during alignment are marked by the black diamonds. (b) After aligning faces to the reference coordinate frame so that the landmarks are comparable. (c) The positions of the aligned landmarks are averaged resulting in the shape average.

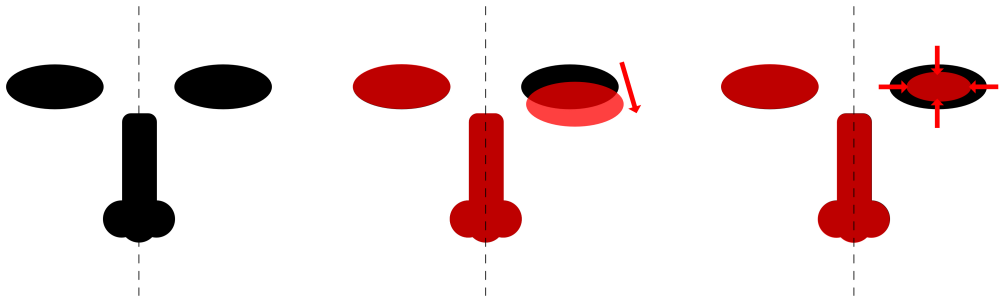
centre of the face, for more consistent alignment (highlighted in Figure 3.3a. Points on the tip of the nose are also ignored, as they are also sensitive to rotation.

The aligned (x, y) position of each corresponding landmark for all faces is averaged as shown in Figure 3.3c.

Asymmetry preservation

We measure differences in relative positioning of aligned landmarks on either side of the face to identify asymmetries. Based on this asymmetry measure, faces are mirrored horizontally to make sure an asymmetrical feature is always present on the same side. This ensures that it is not "averaged out" when creating the shape average.

As facial features vary in size and location, the landmarks are grouped into four face parts for each side: eyes, eyebrows, nose and mouth. The landmarks on the jawline are ignored for the purposes of checking asymmetry, as variations here are mostly due to out-of-plane rotations. Each face part is given two asymmetry scores: one for the difference in shape when compared to its corresponding pair on the other side of the face (shape asymmetry), and one for its relative position to the symmetrical reference constellation (location asymmetry). The two scores measure differences in the shape of facial features (e.g. differently sized eyes) and location differences (e.g. differences in eye positioning), respectively (Figure 3.4).



(a) Reference face **(b)** Location asymmetry **(c)** Shape asymmetry

Figure 3.4: Two different types of asymmetry were considered. The face being considered by the algorithm is shown in red. (a) Symmetric reference constellation (black), (b) location asymmetry considers the relative locations of face parts, which is measured relative to the reference face, and (c) shape asymmetry compares the relative shape of face parts on either side of the face.

The asymmetry is preserved by mirroring each face so that side with the greater sum of asymmetry scores is always on the right side for all faces.

The shape asymmetry of a face part is estimated by the standard deviation of the landmark positions, $\mathbf{x}_i$ of a face part, i, on either side of the face. They are divided by the maximum diameter, d_i, of that face part in the reference constellation to make the measure comparable between differently sized face parts. The contribution of each face part to the total shape asymmetry measure is controlled via a coefficient α_i.

$$A_{\text{shape}} = \sum_{i=1}^{N} \alpha_i \frac{\sigma(\mathbf{x}_i)}{d_i}, \tag{3.1}$$

where $\sigma(\mathbf{x}_i)$ denotes the standard deviation of landmarks for the i-th face part.

The location asymmetry is estimated by taking the average distances between each landmark in a face part and its pair in the reference constellation (see Section 3.2.2). These are also normalised by the maximum diameter of that face part in the reference constellation. Again, the contribution of each face part to the total location asymmetry measure can be controlled via a coefficient α_i.

$$A_{\text{loc}} = \sum_{i=1}^{N} \alpha_i \frac{|\mathbf{x}_i - \mathbf{x}_i^r|}{d_i}, \tag{3.2}$$

where $\mathbf{x}_i^r$ denotes the landmark locations of $i-$th face part in the reference constellation. We consider only one of the asymmetry measures at a time, as they are not directly comparable.

Prior knowledge of where an asymmetry is expected to present can be reflected in choosing larger α-values for the relevant face parts. Face parts that are known not to contain asymmetries for a given syndrome can be ignored by setting the corresponding $\alpha_i = 0$. This reduces the chance that the asymmetry metric is affected by variations in face parts that are known to be irrelevant. We create multiple averages, considering one of shape or location asymmetry for different combinations of α values. The resulting averages are then qualitatively compared to assess whether an asymmetry is present, and the most appropriate average is selected.

In the example in Figure 3.5, we demonstrate asymmetry preservation for images of winking men. As this asymmetry is expected mainly in the eye, we set $\alpha_{\text{nose}} = 0$, $\alpha_{\text{mouth}} = 0$, $\alpha_{\text{eyes}} = 1$, and $\alpha_{\text{eyebrows}} = 1$. The average was created considering only shape asymmetry.

3.2.3 Creating the Appearance Average

The appearance average consists of the average pixel values of each patient's face after warping them onto the shape average.

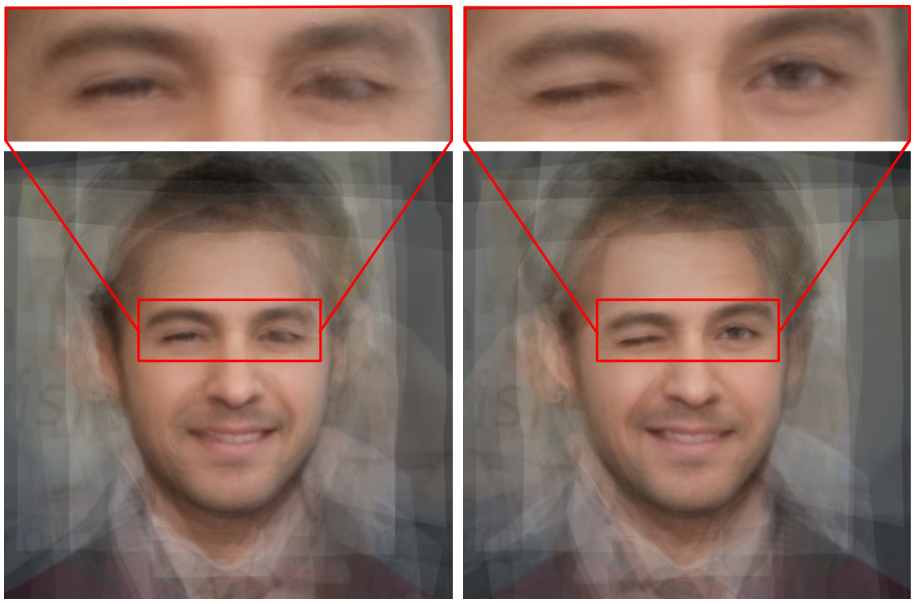
We use the Delaunay triangulation of the positions of facial landmarks to return corresponding regions between faces. Examples of the triangulations are shown in Figure 3.6. Eight additional landmarks are added in a square at a fixed distance around the face, on each corner and halfway point of each of the vertices. This helped make our average faces more realistic by considering the areas around the face, too (Figure 3.6b).

The warping process consists of transforming the image content within each Delaunay triangle of a patient's face constellation to the corresponding triangle in the shape average triangle. This was achieved by an affine transformation. The pixel content for each patient within the transformed triangle is then summed and normalized by the number of images. This process is repeated for each Delaunay triangle in the shape average.

The pixel values of corresponding triangles can vary greatly due to ancestral origin of the patients and variances in illumination. This can become problematic



(a) Input images



(b) Without asymmetry check **(c)** With asymmetry check

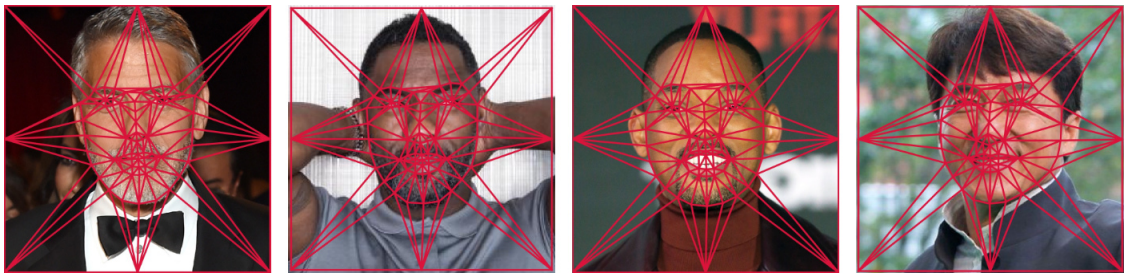
Figure 3.5: Demonstrating left-right facial asymmetry preservation. Images of 10 winking men with the closed eye appearing on either side equally were averaged. (a) A subset of the input images, (b) result without measuring face parts for asymmetry, (c) measuring face parts for shape asymmetry and horizontally flipping to align asymmetry.

for de-identification, as the "odd-one-out" can dominate the average and thus be identified. To mitigate these images from dominating the composite, the intensities of pixel values within the face are normalized to the mean value across all faces in the average. By ignoring the pixel content outside of the face, we make our algorithm insensitive to differences in hair colour and photo backgrounds.

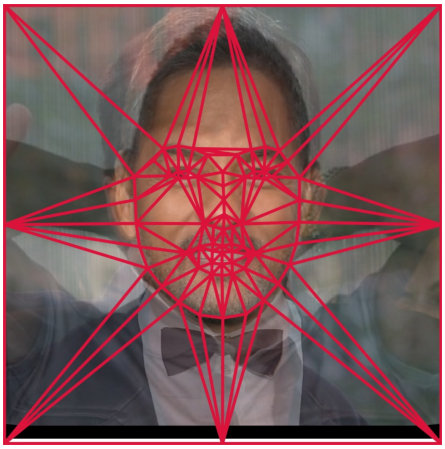
The result of the appearance average is shown in Figure 3.6d.



(a) Annotated faces



(b) Delaunay triangulation of face landmarks including edge points



(c) Delaunay triangulation of shape average



(d) Resulting appearance average

Figure 3.6: (a) Faces are annotated with landmarks and aligned. (b) We add additional landmarks on a square around the face and create the Delaunay triangulation. (c) We then average the pixel content of each triangle after warping it onto the corresponding triangle of the shape average to create the appearance average. The resulting appearance average is shown in (d).

3.3 De-identification

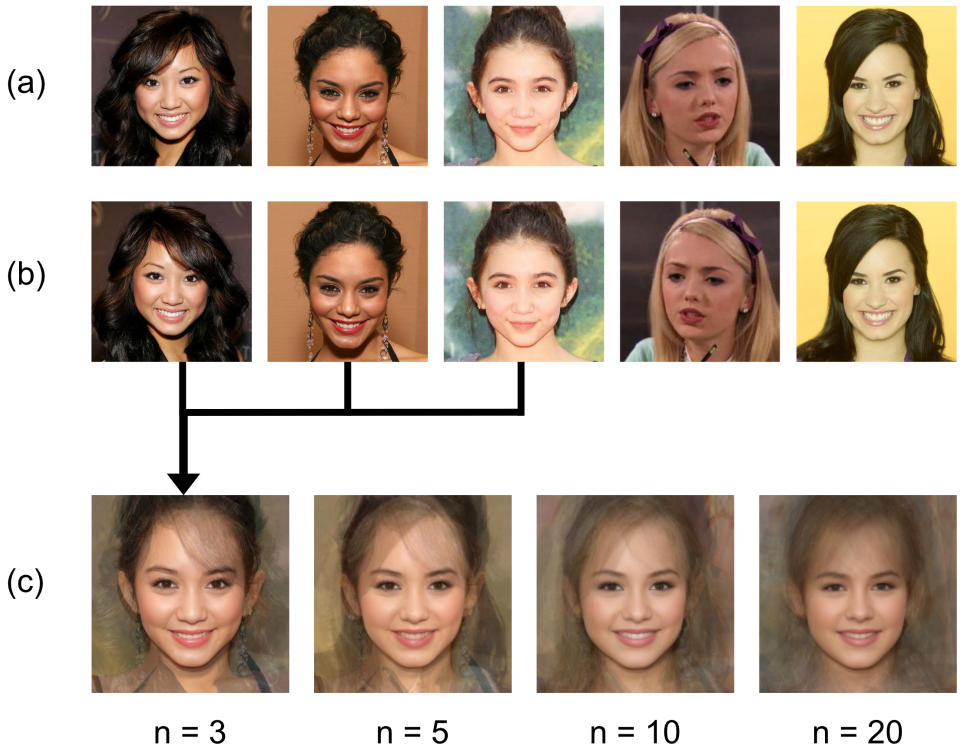


Figure 3.7: De-identification by iteratively adding one image for each identity to the average: First the images (a) are assessed for facial asymmetry and flipped accordingly (b). A number n thereof are selected (from left to right) to create the averages in row (c). Images that were used for $n > 5$ are not shown. We can observe identifiable that constituent individuals become more difficult to identify from $n = 10$ onwards.

As more images are added to the average, the contribution of each constituent image becomes less apparent, until none of them can be identified. Accurate and dense facial landmark detectors result in better averaging of identifiable features. Ferry et al. (2014) use a face annotator which detects 36 landmarks. We increase the number of landmarks to 68, which improve the correspondence between facial features of each face. This results in clearer averages. This allows the creation of average faces for syndromes with fewer images, as identifiable features are averaged out quickly. This is of great importance, as the number of available images of individuals sharing ultra-rare disorders is generally small ($n < 20$).

For cases where we are presented with multiple images for any one patient, to

avoid the average face mesh being biased towards this individual, we compute a shape and appearance average from images of each patient and use the personal average of each patient as inputs to the final average face. Each individual then contributes only once to the final result.

An example of de-identification, when averaging progressively more individuals is shown in Figure 3.7. Clinicians assessed averages created with our methods for results in Deciphering Developmental Disorders Study et al. (2017) and agreed that averages were sufficiently de-identified when consisting of 10-20 individuals.

3.3.1 Quantifying De-identification

We conducted quantitative experiments to assess whether constituent faces could be identified in averages as we increased the number of constituents.

To quantify the level of de-identification, we made use of a Resnet-50 face verification network trained on the VGG-Face 2 dataset (Section 2.4.4). This network encodes faces into a 128-dimensional embedding space in which images of the same individual lie close to each other and different identities lie further apart. We selected images of a 30 random celebrities that were iteratively added to an average. The face embeddings of the resulting averages were compared to each face in a verification set. The verification set, of size N, consisted of 30 images of constituent celebrities (n), which form part of the average, and a number of other "distractor" celebrities ($N - n$). The verification sets are of size $N = 60$ up to $N = 3000$. The verification sets contain one image for each identity.

The distance between the average and each image in the verification set was measured in embedding space and ranked from closest to furthest. We define the identification score $S_N(i)$ as the proportion of constituents (i) that lie amongst the 30 closest images (top-30) to the average. When $S_N(i) = 1$ all constituents are part of the top-30 and the average is deemed perfectly identifiable. When an image is completely de-identified, the constituents and distractors are equally distributed around the average in embedding space, which corresponds to $S^*_N(i) \to \frac{n}{N}$. This number is smaller for larger numbers of distractors in the verification set.

The experiment was repeated 100 times, each time sampling constituents at random from the verification set. The averaged results are shown in Figure 3.8. The identification score decays at a greater rate for the bigger verification sets.

For the hardest case, $N = 60$, where the verification set contains an equal number of constituents and distractors, S_{60} decays as more constituents are added to the average: at $i = 14$ images $S_{60}(14) < 0.55$, which is within 10% of the "perfect" de-identification score $S^*_{60} = 0.5$. The final identification score with all 30 images is $S_{60}(30) = 0.52$. The easiest case, $N = 3000$, where the verification set contains almost 100 times more distractors than constituents, the same score as above is achieved with only two constituents ($S_{3000}(2) = 0.5$) and S_{3000} reaches within 10% of a perfect de-identification score with $i = 12$ constituents.

We suspect the CNN may be "cheating" by making use of image-specific textures in and around the face to measure similarity to an average. Our suspicion comes from the fact that several individuals could not identify which images from the verification set formed part of the average. The network must therefore be able to "see" more than humans would in these averages. When averaging "average-looking" faces with each other, our experiments showed that 5+ images can be sufficient for de-identification. However, creating an average with a very distinctive individual may require more images for de-identification. Generally, qualitative assessments deemed averages to be de-identified after at most 10-12 constituents.

3.4 Results on Clinical Datasets

We made use of a images from the LSITW dataset (Section 2.4.1), the Minerva Image Resource (MIR) (Section 2.4.1), and images sourced from online publications via PubMed (Section 2.4.1). Averages were only created for syndromes that were shared by more than 20 patients and that resulted in satisfactorily de-identified averages. The averages are shown in Figure 3.9 and the sources of images used in each average are detailed in Table 3.1.

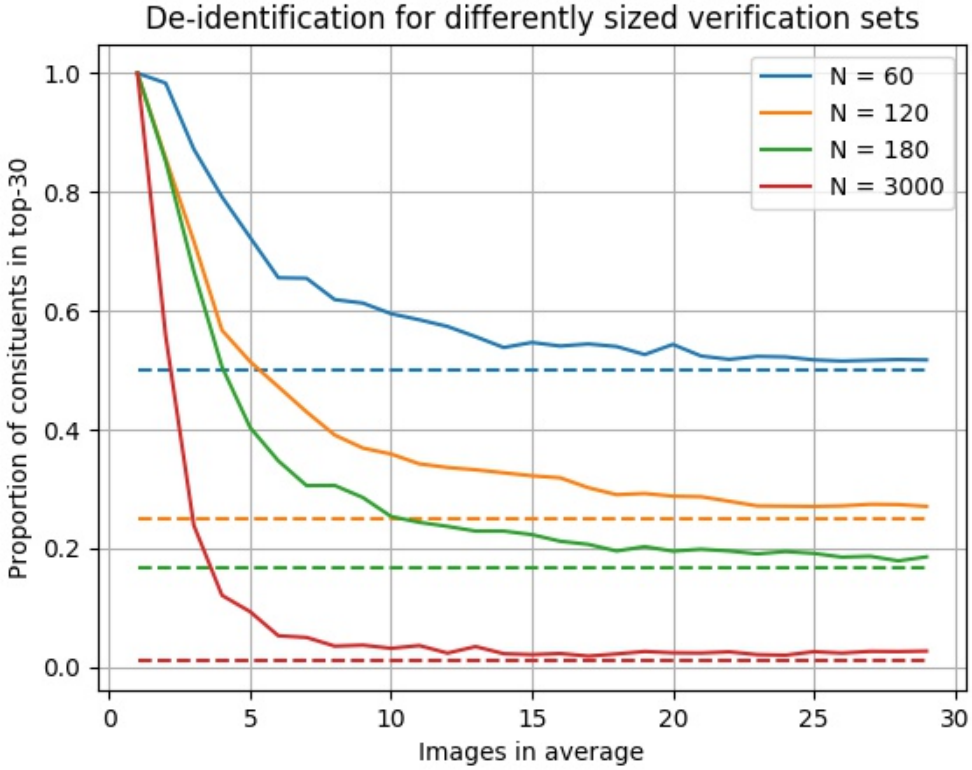


Figure 3.8: Identification score for averages of increasing number of individuals. A score of $S(i) = 1$ equates to a perfectly identifiable average, when the feature representations of constituents are amongst the top-30 most similar embeddings. As $S(i)$ decays, the averages become less identifiable. The dashed lines show the value of $S_N^* = \frac{n}{N}$ for each verification set at which the constituents are randomly distributed around the average amongst distractors and thus the average is perfectly de-identified.

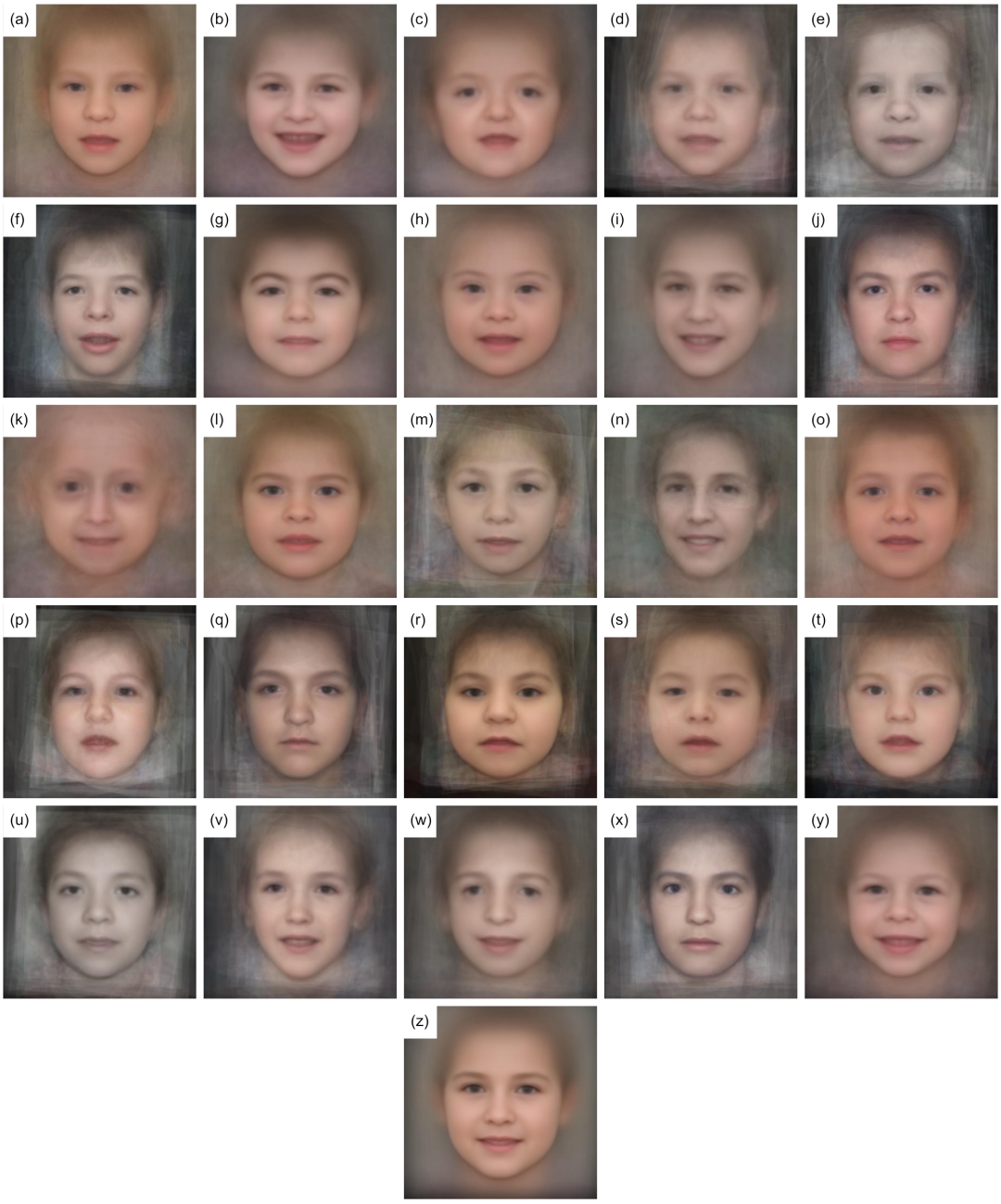


Figure 3.9: Average faces that were generated using methods detailed in Section 3.2. The source of the images and the number of patients in each composite are listed in Table 3.1.

Table 3.1: Reference table with the image source for each average in Figure 3.9.

Caption	Syndrome	Number of Patients	Source
a	17q21-31 microdeletion syndrome	49	MIR
b	Angelman syndrome	205	LSITW
c	Apert syndrome	203	LSITW
d	B-RAF proto-oncogene serinethreonine kinasa syndrome	36	MIR
e	Cardiofaciocutaneos syndrome	25	MIR
f	Coffin-Lowry syndrome	38	MIR
g	Cornelia de Lange syndrome	509	LSITW, MIR
h	Down syndrome	192	LSITW
i	Fragile X syndrome	209	LSITW, MIR, Pubmed
j	Gorlin syndrome	39	MIR
k	Hutchinson Progeria syndrome	158	LSITW, MIR
l	Kleefstra syndrome	63	MIR
m	Mandibulofacial dysostosis	24	MIR
n	Marfan syndrome	37	LSITW, PubMed
o	Noonan syndrome	64	MIR
p	Orofaciodigital syndrome	21	MIR
q	Otopalatodigital syndrome	26	MIR
r	PGAP3	25	MIR
s	PIGA	25	MIR
t	PIGV	24	MIR
u	SOS RasRac guanine nucleotide exchange factor 1	30	MIR
v	Sotos syndrome	46	MIR
w	Treacher-Collins syndrome	103	LSITW
x	Waardenburg syndrome	41	MIR
y	Williams syndrome	232	LSITW
z	Healthy controls	1515	LSITW

3.5 Creating Transitions Between Faces

Animations of transitions between faces can help highlight the differences between them.

We create transitions by linearly interpolating between pairs of aligned landmarks, X_1 and X_2, of both faces creating intermediate faces (Equation 3.3). The pixel values for both faces, A_1 and A_2, are blended in the same ratio as the shape interpolation, as shown in equation 3.4. The parameter γ denotes the ratio of the interpolation between both faces, where $\gamma = 0$ is the beginning of the interpolation (i.e. only the first image), and $\gamma = 1$ is the end of the interpolation. We animate the transitions, or "morphs", by creating videos with each intermediate step representing a frame of the video.

Some examples are shown in Figure 3.10. In each case, a healthy control average is used as the starting image and it is gradually "morphed" into a syndrome average face. These syndrome averages in Figure 3.10 are created using the images from the LSITW dataset (Section 2.4.1). Videos of these transitions available via links given in Table 3.2.

$$X(\gamma) = \gamma X_1 + (1 - \gamma)X_2, \tag{3.3}$$

$$A(\gamma) = \gamma A_1 + (1 - \gamma)A_2. \tag{3.4}$$

3.6 Facial Phenotype Caricatures

We take a similar approach as in Section 3.5 to create new face shapes that exaggerate syndrome phenotypes. However, instead of creating intermediate faces, by interpolating between a healthy control and a syndrome average, we extrapolate beyond the syndrome average. These exaggerated caricatures further highlight the syndrome characteristic deviation from a healthy face.

We calculate the deformation of the face mesh from a healthy control average to a syndrome average using the vectors connecting the line between a pair of corresponding face landmarks. An exaggerated average face shape is created by

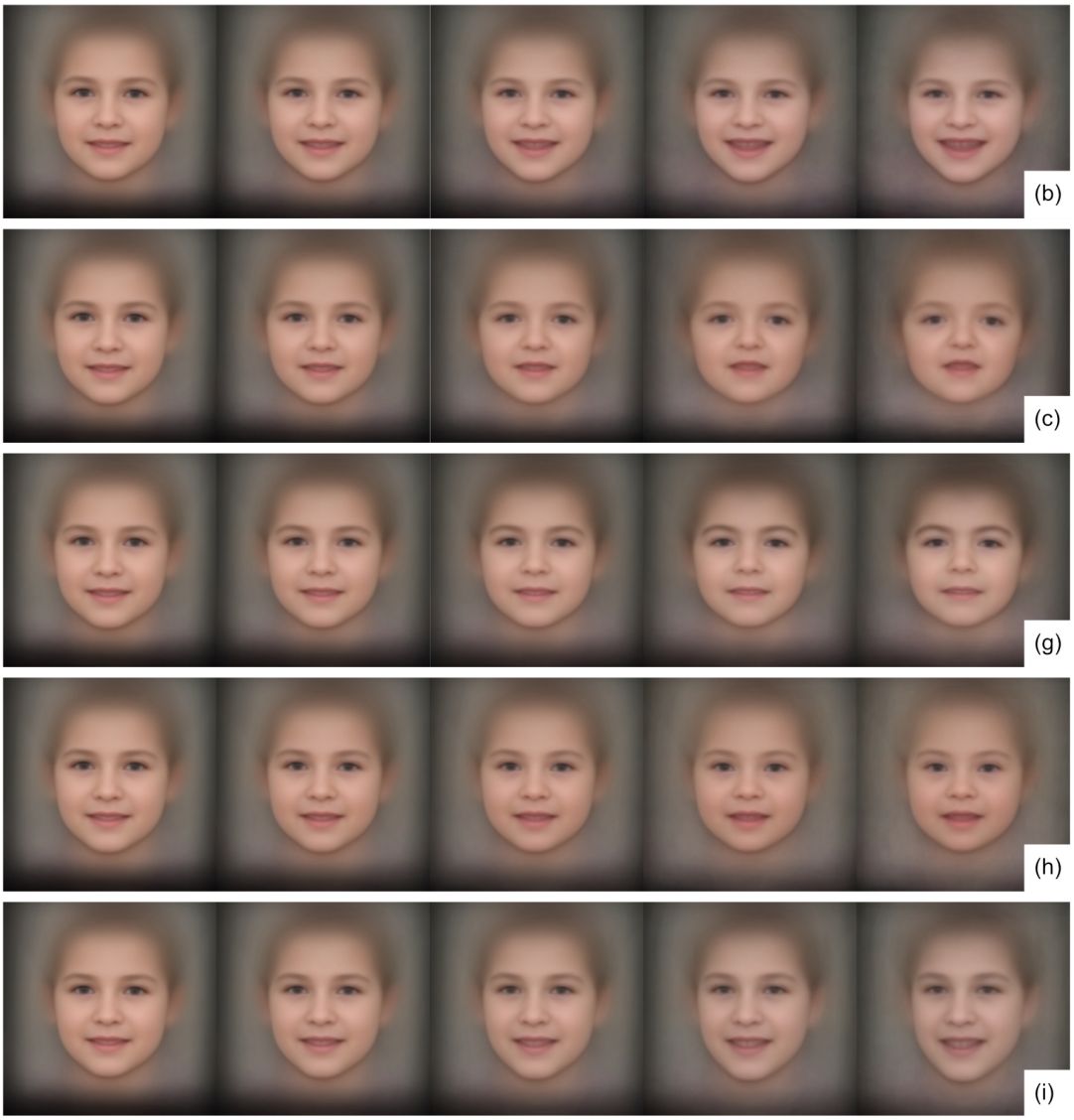


Figure 3.10: Transitioning from a healthy control averages to average faces of syndromes. The image source and name of disorders are given in Table 3.1. The videos of these and more syndromes are published at the links in Table 3.2

moving further along these vectors, beyond the syndrome average. Figure 3.11 shows caricature version of the same set of faces as in Figure 3.9. Creating videos of the transition from healthy controls to a syndrome average and beyond it to a caricature, further highlights the facial phenotype. A limitation of this approach is that faces can explode or collapse if extrapolated too far beyond the syndrome average.

Table 3.2: Videos of transitions from healthy control averages to syndrome averages

17q21-31 microdeletion syndrome	https://youtu.be/Xg6qR75qmWc
Angelman syndrome	https://youtu.be/OUsuGlS_IXU
Apert syndrome	https://youtu.be/YnNN15Z0TRM
B-RAF proto-oncogene serinethreonine kinasa syndrome	https://youtu.be/XU6f1-RFQs0
Cardifaciocutaneous syndrome	https://youtu.be/que-SO2PT6o
Coffin Lowry syndrome	https://youtu.be/rMrRo2cDgjI
Cornelia de Lange syndrome	https://youtu.be/6eiF7bbDvPQ
Down syndrome	https://youtu.be/znpDqmrkgSw
FragileX syndrome	https://youtu.be/Koc4MJ4XZXc
Gorlin syndrome	https://youtu.be/1TQe5lNJ2_o
Hutchinson Progeria syndrome	https://youtu.be/N3yGxVF5sfE
Kleefstra syndrome	https://youtu.be/xURPg1DeycM
Mandibulofacial dysostosis	https://youtu.be/5B0WI8mzOvM
Marfan syndrome	https://youtu.be/SZrszUu-NiE
Noonan syndrome	https://youtu.be/B-XN4pyz1oM
Orofaciodigital syndrome	https://youtu.be/sbja2rmR-9s
Otopalatodigital syndrome	https://youtu.be/NlDiHEYOs90
PGAP3	https://youtu.be/Fb0ekbGaDfk
PIGA	https://youtu.be/7ZFewFhprLg
PIGV	https://youtu.be/RDaMVxD1oFI
SOS RasRac guanine nucleotide exchange factor 1	https://youtu.be/OGbFsgl9O0k
Sotos syndrome	https://youtu.be/g031pVQ6lnI
Treacher-Collins syndrome	https://youtu.be/UIbwPRtmXvI
Waardenburg syndrome	https://youtu.be/eIQvTFsNdIA
Williams syndrome	https://youtu.be/gtpLjFncvBQ



Figure 3.11: Caricature average faces that were generated by exaggerating syndrome faces by 50%. The source of the images and the number of patients in each composite are listed in Table 3.1.

3.7 Conclusions

We presented our methods for creating realistic average faces that preserve syndrome phenotypes, whilst de-identifying constituent individuals. Qualitative inspection of averages consisting of faces that were not fully frontal showed that our alignment algorithm resulted in clearer averages. Additionally, we showed that our average faces preserve left-right facial asymmetry. We also presented methods for creating exaggerated averages and transitions between healthy controls and syndromes averages. These can be used to further highlight the clinically relevant phenotype.

The results of our algorithm have been published in several clinical publications (Reijnders et al., 2018a), (Reijnders et al., 2018b), (Bengani et al., 2017), (Deciphering Developmental Disorders Study et al., 2017).

4

Classification of Genetic Disorders

Contents

4.1 Classifying Syndromes with Engineered-Feature Based Representations 57
4.1.1 Previous Work 58
4.1.2 Improving the Shape Descriptor 60
4.1.3 Improving the Appearance Descriptor 63
4.1.4 Classification Experiments 66
4.1.5 Results and Discussion 67
4.2 Classifying Syndromes using Deep Representations . . 71
4.2.1 Networks and Pre-training 72
4.2.2 Fine-tuning for Syndrome Classification 73
4.2.3 Part-based Models 74
4.2.4 Image Augmentation 74
4.2.5 Results and Discussion 76
4.3 Conclusions 78

This chapter discusses the classification of genetic disorders from facial photographs of individuals in uncontrolled settings, in contrast to previous work, that required fixed conditions (Section 2.1.1). We make use of the LSITW dataset that is described in Section 2.4.1 to demonstrate these algorithms. The images contain a great amount of variety in pose, lighting, expression, age, ancestral origin, and background. It is therefore imperative to find robust representations that capture the clinically relevant aspects of facial variation between syndromes.

We extend prior work by Ferry et al. (2014), who developed engineered-feature based representations of shape and appearance to classify the 8 syndromes and controls in the LSITW dataset. These representations rely on local image descriptors, that are extracted from around defined facial landmarks that are detected automatically. The use of local feature descriptors gives rise to two main challenges. First, the locations around which the features are computed need to be accurate and consistent across images. In an automated setting, these locations must be found with the help of a landmark annotation algorithm. And secondly, the features must be informative and robust to noise sources.

We summarise the approach taken by Ferry et al. (2014) in Section 4.1 and discuss its limitations. We then present our methods for more accurate landmark detection and also present an alternative, dense appearance feature description. These feature representations are used to conduct classification using linear support vector machines (SMV). We also investigate the use of CNNs to create deep representations for image classification (Section 4.2). CNNs usually require training datasets that are far larger than the LSITW dataset. We make use pre-training, fine-tuning and image augmentation to overcome this limitation. Lastly, we explore the use of part-based models, where separate models are trained to conduct inference on different parts of the face.

This chapter made use of software from colleagues to generate SDM models and Fisher Vectors.

4.1 Classifying Syndromes with Engineered-Feature Based Representations

Engineered-feature based computer vision consists of two distinct steps: 1) feature engineering, finding a way to translate pixel-values into a feature representation of an image, and 2) inference, using a decision framework that can interpret the features to conduct inference.

Using pixel values alone to describe an image, e.g. of a face, would not result in a robust feature representation, as changes in expression, pose, lighting, and scale

significantly affect the pixel values and therefore the representation. Two instances of the same person shown with different expressions would then be unrecognisable for classification. A good feature representation, should be invariant to these noise sources. Features can be extracted either at points of interest in the image (e.g. local features, around facial features, such as the nose) or densely (extracting features at many locations across the entire image). They are then combined, either by simply concatenating them, or encoding them using a statistical model.

Ferry et al. (2014) created two face representations that are based on local features: a shape descriptor and an appearance descriptor (Section 4.1.1). These were used to train linear SVMs for classification. Both of these descriptors have drawbacks, that lead to relatively inaccurate descriptors, which motivated us to investigate better descriptors:

1. A more accurate landmark annotation algorithm to improve the shape descriptor (Section 4.1.2)

2. An improved appearance descriptor using Fisher vectors (Section 4.1.3).

4.1.1 Previous Work

This section describes the shape and appearance descriptors that were used individually and jointly, by means of data fusion, for the facial detection of genetic disorders by Ferry et al. (2014).

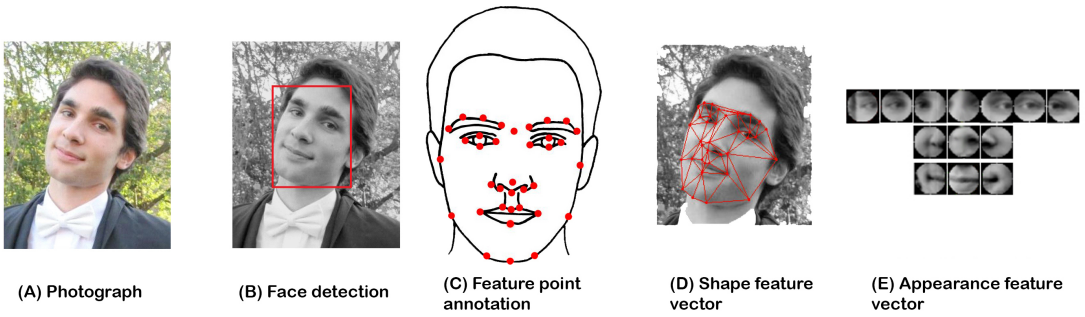


Figure 4.1: Overview of the shape and appearance descriptor pipeline in Ferry et al. (2014). **(A-B)** First the face is detected and converted to gray-scale. **(C-D)** The pair-wise distances between 36 automatically annotated landmarks comprise the shape descriptor. **(E)** The pixel values of circular patches around 9 automatically detected landmarks comprise the appearance descriptor.

Shape Descriptor

The shape descriptor of Ferry et al. (2014) consists of 630 pairwise distances between 36 automatically annotated landmarks. The landmarks are located across the supra-orbital ridge (8 points), the eyes (mid points of the eyelids and eye canthi, 8 points), nose (nasion, tip, ala, subnasale and outer nares, 7 points), mouth (vermilion border lateral and vertical midpoints, 6 points) and the jaw (zygoma mandibular border, gonion, mental protruberance and chin midpoint, 7 points) (Figure 4.1C).

The locations of these landmarks were predicted using a consensus of exemplars model (Belhumeur et al., 2011), which reconciles the local estimates of the landmarks with a global model of their expected configuration on the face. This algorithm was initialised relative to the location of a face within an image, which was detected using an adapted implementation from Viola and Jones (2004).

To make the feature representation comparable across images of different sizes, each face's landmarks were registered via the Procrustes transformation to the landmarks of an average face model and normalised by dividing by the pair-wise distances in the average face model.

The locations of the the predicted facial landmarks using this annotation algorithm were often inaccurate along the jaw-line (Figure 4.2a), which lead to inaccurate shape feature representations. This motivated us to investigate methods to improve automatic landmark annotation (Section 4.1.2).

Appearance Descriptor

The appearance feature vector was based upon the gray-scale pixel content of cropped patches around 13 facial landmarks on the inner face (Figure 4.1E). A constellation of 9 of these landmarks was detected automatically using a facial landmark module (Everingham et al., 2009). These were used to generate four additional points: left and right center of the eye, the nasion (the point between the eyes), and center of mouth. These were used to register the face, via an affine transformation to a canonical set of corresponding points, to account for scale and orientation variation between faces.

The appearance descriptor is sensitive to illumination as it is based on pixel values. It also fails to describe the complete face, most noticeably the jaw-line. To mitigate these drawbacks, we investigated dense appearance description of the complete face with an illumination invariant features (Section 4.1.3).

4.1.2 Improving the Shape Descriptor

We developed an improved facial landmark predictor to create a more accurate shape descriptor than the one described in Section 4.1.1. This was achieved using the Supervised Descent Method (SDM) (Xiong and De la Torre, 2013). The objective of SDM is to learn a generic descent map, that, given an initial estimate for the location face landmarks, moves them to the correct location. This method has been used successfully in a number of face annotation and tracking problems (Liu et al., 2015; Xiong and De la Torre, 2015).

Finding the optimal landmark locations is posed as a non-linear optimization problem. Given an initial estimate of landmark locations $\mathbf{x}_0$, the aim is to find the unknown, true location of the landmarks $\mathbf{x}_*$. These locations are described by a local feature $\phi_{\mathbf{x}}$, in our case the SIFT descriptor, around a landmark, $\mathbf{x}$. Finding the correct landmark locations can be formulated as the minimization of the following function:

$$f(\mathbf{x}_0 + \Delta\mathbf{x}) = \|\phi_{\mathbf{x}} - \phi_{\mathbf{x}_*}\|_2^2 \tag{4.1}$$

Using Newton's method, the update step for this minimization requires computation of the Jacobian and Hessian matrices (Equation 4.2). Since the SIFT descriptor, $\phi_{\mathbf{x}}$, is not differentiable, these quantities need to be numerically approximated, which is computationally expensive (Xiong and De la Torre, 2013).

To overcome this limitation, Xiong and De la Torre (2014) reformulated the Newton update step

$$\Delta\mathbf{x}_{n+1} = -2\mathbf{H}^{-1}\mathbf{J}^\top(\phi_{\mathbf{x}_t} - \phi_{\mathbf{x}_*}), \tag{4.2}$$

as a N-step iterative procedure of average descent directions $\mathbf{R}_t$ and biases $\mathbf{b}_t$, where $t = 1, ..., T$, that produce a set of updates that iterate from $\mathbf{x}_0$ to $\mathbf{x}_*$ (Equation 4.3).

$$\Delta \mathbf{x}_{t+1} = \mathbf{R}_t \phi_{\mathbf{x}_t} + \mathbf{b}_t, \tag{4.3}$$

where $\mathbf{J}$ is the Jacobian and $\mathbf{H}$ is the Hessian matrix of the feature descriptors. Given a training dataset $\mathcal{D} = \{(d^1, \mathbf{x}_*^1), (d^2, \mathbf{x}_*^2), \ldots\}$ that contains images $\mathbf{d}^i$ and the known optimal landmark locations, $\mathbf{x}_*^i$, the directions and biases $\mathbf{R}_t$ and $\mathbf{b}_t$ are found as follows:

$$\arg\min_{\mathbf{R}_t, \mathbf{b}_t} \sum_{\mathbf{d}^i} \sum_{\mathbf{x}_t^i} \|\Delta \mathbf{x}_*^i - \mathbf{R}_t \phi_{\mathbf{x}_t^i} - \mathbf{b}_t\|^2. \tag{4.4}$$

We used many possible initialisations of $\mathbf{x}_0^i$ for each image to estimate the first set of descent directions and biases $\mathbf{R}_0$, $\mathbf{b}_0$. Xiong and De la Torre (2013) showed experimentally that learning 4 to 5 update steps were sufficient for convergence. At test time, the update rule in Equation 4.3 is applied sequentially for each t starting from a random initialisation. The landmark locations after the last update step are the predicted locations of the landmarks.

SDM Annotator Training

This section provides more details about how the SDM landmark annotator was trained.

This algorithm was implemented with the same 36 landmark locations to create a comparable shape feature vector as the one used by Ferry et al. (2014) (Section 4.1.1). The LSITW dataset, for which ground truth landmarks are available, was used for training and evaluating the annotation model. The images of healthy controls were flipped horizontally and along with the originals used for training, resulting in 3,030 training images. The remaining images in LSITW were used as the test dataset.

First, faces and their orientations were found by rotating images in 15°-increments from -45°to 45°. We selected the rotation that returned the highest detector score returned by the face detector from Viola and Jones (2004). The upright face was saved along with the locations of the corners of the face detection. The faces

were then scaled so that the face detection box height was 140 pixels. The facial landmarks were initialised relative to the face detection box and consist of the average ground truth locations of all landmarks in the training dataset. Each image's landmarks were randomly perturbed by additive Gaussian noise to create 10 further initialisations for training the descent directions.

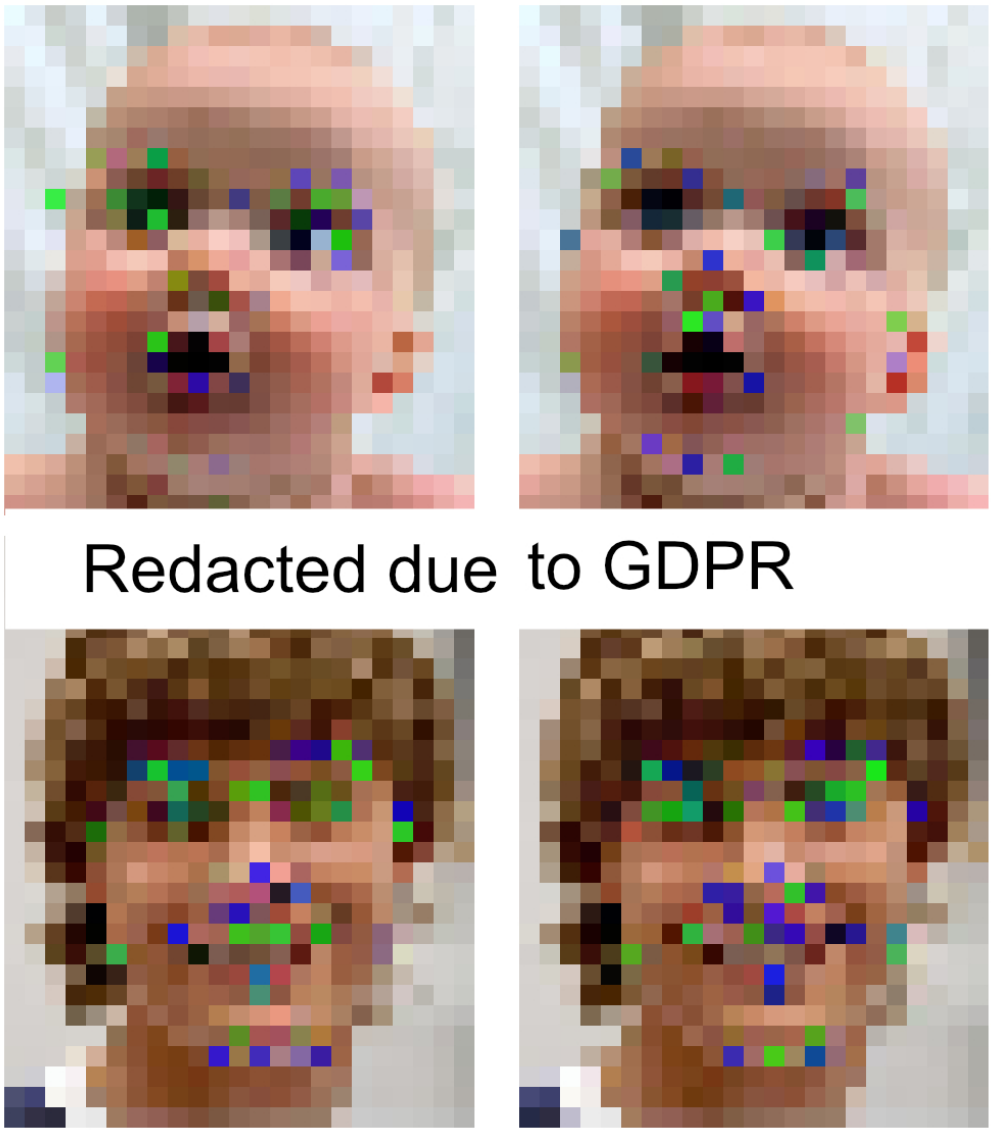
We extracted SIFT features of size 24x24 pixels around each landmark and it's optimal position. These were dimensionality-reduced by PCA, preserving 98% of the signal. The projected features were used as inputs to Equation 4.4. The training procedure was conducted for 4 update steps.

SDM Annotator Evaluation

We compared the accuracy of the SDM annotator and the annotator used by Ferry et al. (2014) (Section 4.1.1). This was done by comparing the distances between each method's predictions on the LSITW test dataset and the corresponding ground truth locations.

The errors were measured as the Euclidean distance between predicted annotation locations and the ground truth. To make the errors comparable across training images, they were normalised by the ground truth inter-ocular distance. The average normalised error using methods from Section 4.1.1 is 0.22. When using our SDM model, this reduces to 0.11, a 50% reduction in error. An example of the performance of the two annotation algorithms is shown in Figure 4.2. The improvements were most noticeable along the jaw-line of individuals with non-neutral facial expressions, e.g. when the subject is laughing. Also, significant improvements were observed along for describing eyebrows, especially when low-resolution images were being annotated.

This increased accuracy should also result in more accurate face descriptors. The shape descriptor calculated with the predicted landmark positions using SDM be referred to as Shape (SDM).



(a) Ferry et al (Section 4.1.1) **(b)** SDM

Figure 4.2: Comparison of (a) original face landmarks for shape feature vector computation from Ferry et al. (2014) (Section 4.1.1) and (b) improved landmarks using SDM 4.1.2 (green crosses). In each case, manually annotated ground truth points (blue circles) are shown. Note in particular the improvement in landmarks around the chin.

4.1.3 Improving the Appearance Descriptor

The appearance descriptor in Section 4.1.1 only describes a fraction of the face. One possible improvement would be to extract features densely, across the entire face, thus capturing more information. We make use of SIFT descriptors (Lowe, 2004), which extract gradient information from local patch in an image and more robust

than simply using pixel values. Extracting these descriptors across the entire face and concatenating them, however, would result in a very large appearance descriptor. We therefore make use of Fisher Vectors, as implemented by Simonyan et al. (2013), to encode the dense SIFT descriptors into a feature descriptor of manageable size.

The Fisher Vector (FV) is an improvement upon the 'Bag of Visual Words' encoding technique of learning 'dictionaries' of standard features and assigning local descriptors (SIFT in this case) to them (Sivic and Zisserman, 2003). A vocabulary of visual words is generated by creating a Gaussian mixture model (GMM) with K-centres of SIFT vectors extracted from the training dataset. The centres of the clusters are defined as words. An image is then described by finding the correspondences between its SIFT features and the GMM (Sanchez et al., 2013).

Let $L = (x_1, x_2, ..., x_N)$ be the N dense SIFT descriptors describing the image. Let $\Theta = (\mu_k, \Sigma_k, \pi_k : k = 1, 2, ..., K)$ be the means, covariance matrices and weights, respectively, of the GMM with K centres, fitting the distribution to the descriptors. Before the GMM is trained on the SIFT descriptors, the descriptors are decorrelated by PCA. This allows the use of diagonal covariances when training the GMM. The first and second order differences between the dense features and each of the GMM centres is then given in Equations 4.5 and 4.6, respectively.

$$\Theta_k^{(1)} = \frac{1}{N\sqrt{\pi_k}} \sum_{p=1}^{N} \alpha_p(k) \left(\frac{x_p - \mu_k}{\sigma_k} \right), \tag{4.5}$$

$$\Theta_k^{(2)} = \frac{1}{N\sqrt{2\pi_k}} \sum_{p=1}^{N} \alpha_p(k) \left(\frac{(x_p - \mu_k)^2}{\sigma_k^2} - 1 \right), \tag{4.6}$$

where $\alpha_p(k)$ is the soft assignment weight of the p-th feature x_p to the k-th Gaussian, and σ_k is the standard deviation for the $k-$th component of the GMM. The FV, Φ, is constructed by stacking the differences for all Gaussians (Equation 4.7). The FV is post-processed as described in the "Improved FV" encoding (Perronnin et al., 2010) by signed root-squaring and L_2 normalisation.

$$\Phi = \left[\Theta_1^{(1)}, \Theta_1^{(2)}, ..., \Theta_K^{(1)}, \Theta_K^{(2)} \right] \tag{4.7}$$

Implementating Fisher Vector Faces

This section details the implementation procedure for extracting FV descriptors from faces.

Facial landmarks were automatically detected using our SDM model from Section 4.1.2 and used in the Procrustes transform to transform faces to a canonical frame. A 160×125 face region containing the landmarks was extracted for further processing. SIFT descriptors were found for 24×24 pixels patches with a stride of one pixel at five scales. We used PCA to reduce the SIFT dimensionality from $d = 128$ to $d = 64$, we used $K = 512$ Gaussians in the GMM. This amounted to a FV descriptor of dimensionality $2Kd = 65536$. An overview of this method is shown in Figure 4.3.

The FV descriptor now effectively encodes the feature space of the image, but has no information about the spatial distribution. To include information about the spatial location of each of the descriptors, the coordinates of the center of the patch (x, y) were appended to each SIFT after PCA: $[S_{xy}; \frac{x}{w} - \frac{1}{2}; \frac{y}{h} - \frac{1}{2}]$, where S_{xy},w, h are the SIFT after PCA at location (x, y), the width, and height of a patch, respectively. The final FV dimensionality as a result was $2K(d + 2) = 67584$.

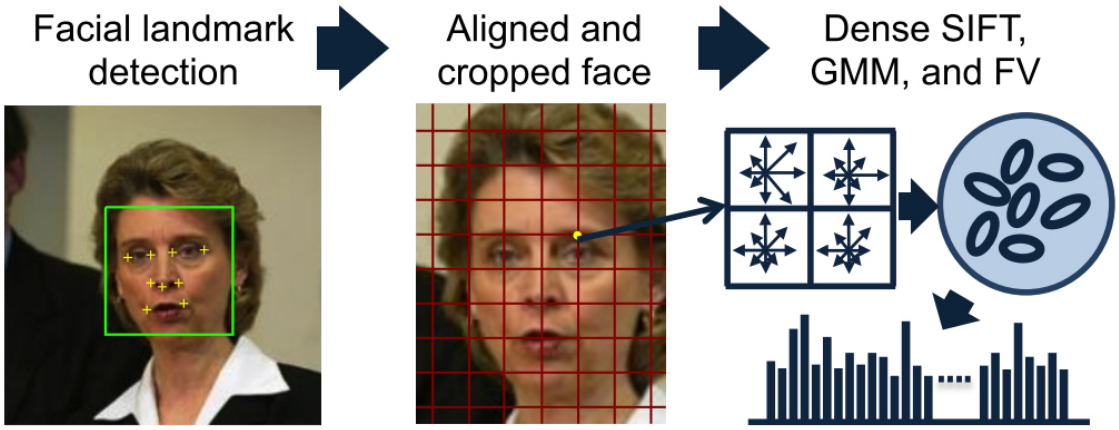


Figure 4.3: Overview of FV description of an image. First the face was detected and landmarks found using Everingham et al. (2009). These were used to align and crop the face. The dense SIFT was determined for this image and decorrelated using PCA, and used as the input to a GMM. The FV was constructed by stacking the differences for all Gaussians of the GMM. Figure adapted from Simonyan et al. (2013)

4.1.4 Classification Experiments

The following contains an extensive evaluation of face descriptors introduced in this section in a variety of different classification scenarios. We make use of the LSITW dataset, described in Section 2.4.1. The training dataset consists of 80% of images from each class, and the test dataset consists of the remaining 20%. There are a total of 8 syndrome classes and a healthy control class.

We investigate three tasks: binary classification of syndromes (multiple two-class problems of syndromes against each other), forced classification of syndromes (an 8-class problem), and the detection of any disorder being present (healthy patients against all syndromes, a two-class problem). In all cases, the performance was measured by the average classification accuracy on the test dataset. This is the proportion of correctly classified images for each class and can be computed as the average of the diagonal of the row normalized confusion matrix. For a random classifier, this equates to a 50% accuracy for a two class problem, and 13% for an 8 syndrome forced classification case. We investigated the performance of each descriptor individually as well as a fusion of shape and appearance descriptors to make use of both sources of information simultaneously.

We used a support vector machine (SVM) with a linear kernel as the classifier. Optimal regularization parameters were found using 5-fold cross-validation for each classifier on the training data. Decision values were trained to give equal error rates (number of false positives = number of false negatives). We report the average classification accuracies on the test dataset.

Additionally, to augment the datasets, both original and horizontal mirror versions of each image were used. The corresponding mirrored image for each image was added to the same set, training or test, as its original.

Experiment 1: Binary Classification

Classifiers were trained for each of the 28 pairwise combinations of 8 syndromes. For each case, class sizes were balanced to contain an equal number of images

by sub-sampling the majority class to avoid bias towards syndromes for which more images were available.

Experiment 2: Forced Classification of Syndromes

This experiment consists of trying to identify which of the 8 syndromes a query image belongs to.

We trained a one-vs-rest classifier for each syndrome, amounting to 8 classifiers, each classifying whether or not an image contains that syndrome. Every test image is presented to all classifiers $[C_1, .., C_8]$ and the decision value of each classifier was used as a probabilistic estimate P_i for the test image to belong to class i. The test image was then assigned to the class with the highest probabilistic estimate P_i.

Experiment 3: Binary Classification of Healthy vs Syndromes

To distinguish between healthy controls and people with one the 8 syndromes in the database, a two-class "healthy vs syndrome" classification case was investigated. In this analysis, the actual syndrome is of no concern, thus a classifier was trained using two classes, one with control images, and one that contained images from all syndromes. The classification accuracies were balanced by classes, so that the larger syndrome classes do not outweigh the smaller ones.

Early vs Late Data Fusion

We investigated data fusion to use shape and appearance features jointly using early and late fusion. Early fusion consists of stacking individually L2-normalised feature vectors and training a single SVM classifier for the features. For late fusion the decision values from separate SVM classifiers, each trained on a single feature, are summed. Features that were combined in this way are denoted with a prefix denoting the fusion method (early: EF, or late: LF).

4.1.5 Results and Discussion

This section contains classification accuracies for each experiment described in Section 4.1.4 for the different feature descriptors:

- **Shape** — Shape descriptor, as proposed by Ferry et al. (2014). (Section 4.1.1).
- **Appearance** — Appearance descriptor, as proposed by Ferry et al. (2014). (Section 4.1.1).
- **Shape SDM** — Shape descriptor using SDM (Section 4.1.2).
- **Shape GT** — Shape descriptor using the ground truth landmark locations. This represents the best-case shape feature, which provides a theoretical ceiling to compare annotations from the automated algorithms to. (Section 4.1.2).
- **FV** — Appearance descriptor using FV (Section 4.1.3).

A summary of all features for each experiment is given in table 4.1. The average binary 1v1 accuracy was calculated by averaging the classification accuracies for all 28 binary classifiers. The average forced classification accuracy corresponds to the mean-per-class accuracy.

First the two shape descriptors are compared, followed by the appearance descriptors. Then we present data fusion results followed by a short discussion about the syndromes.

Shape vs Shape SDM

Shape descriptors calculated using the SDM method significantly outperform the shape descriptors used by Ferry et al (Section 4.1.1) in all experiments. This difference is most apparent for the forced classification case, where Shape SDM improves the average classification accuracy by 36.2% from 45% (Shape) to 81.2% (Shape SDM). Similarly, the binary 1v1 and syndromes vs control accuracies are significantly better when using SDM as the basis for the shape descriptor (+12.7% and +14.2%, respectively). This highlights the importance of accurate landmark localisation for shape description. When compared to shape features based on the ground truth annotations, it can be seen that Shape SDM gets within 0.4% and 0.1% of the accuracy that would have been achieved if all landmark location

Table 4.1: Summary table of classification accuracies (percent) on test data using engineered-feature based feature representations. The binary 1v1 average accuracy corresponds to the average of each 1v1 classifier accuracy. The forced classification average accuracy is the average of the diagonal of a normalised confusion matrix.

	Binary 1v1	Forced Class.	Syndromes vs Control
Shape GT	93.9	84.9	88.7
Shape	80.8	45.0	74.4
Shape SDM	93.5	81.2	88.6
Appearance	92.2	74.9	84.0
FV	94.8	85.5	88.5
(EF) Shape + Appearance	92.9	76.3	83.4
(LF) Shape + Appearance	91.3	64.5	82.2
(EF) Shape SDM + FV	95.3	86.7	88.9
(LF) Shape SDM + FV	**95.9**	**88.4**	**91.3**

were found perfectly. The difference was more pronounced for the forced case (difference of 3.7%), which shows that there is still a potential benefit in pursuing even more accurate annotation algorithms.

Appearance vs FV

FV features significantly outperform Appearance. In the simpler task of binary 1v1 classification, the average accuracies improved by 2.6% (92.2% to 94.8%). In the more challenging task of forced classification, the improvement was more pronounced: +10.6% (74.9% to 85.5%). When trying to distinguish syndromes from controls, FV (88.5%) also outperformed Appearance (84.0%).

FV is based on dense SIFT, which is invariant to illumination and describes all of the image. In contrast, Appearance relies on localised pixel values, which can be affected by illumination and, most noticeably, do not describe the jaw-line.

Data Fusion

Early and late fusion perform very differently for the two sets of descriptors:

When combining Shape and Appearance, late fusion causes accuracies to drop even lower than when considering Appearance alone. For early fusion

of Shape+Appearance the accuracies improve for the binary 1v1 and forced classification experiments, but deteriorate for syndromes vs controls. This is most likely due to the comparatively poorer quality Shape descriptor. For early fusion the SVM classifier may learn to ignore most of the noise introduced by the Shape descriptor, late fusion weights the outputs of both Appearance and Shape classifiers equally.

Both fusion approaches improve classification scores for all experiments when combining Shape SDM and FV. Late fusion performs better than early fusion in all experiments. The best results in all experiments are achieved for late fusion of Shape SDM + FV: They improve upon FV, the best non-fusion descriptor, by 1.1% (binary 1v1), 2.9% (forced classification) and 2.8% (syndromes vs control). Detailed results for binary 1v1 classification and forced classification are given in figures 4.4 and 4.5, respectively.

Syndrome Discussion

The Angelman syndrome is the most difficult syndrome to classify (76% accuracy, forced classification), despite being the second largest class by volume (205 images). By contrast, the smallest class, Hutchinson-Progeria (78 images), performs far better in both 1v1 and forced experiments. Fragile X (164 images) is the second hardest syndromes to classify with a classification accuracy of 8.8% lower than the average in the forced classification case. The remaining syndromes are significantly easier to predict with 91+% classification accuracies for forced classification.

These results concur with qualitative observations, as some syndromes manifest more prominently in the face, which can be seen in the averages in Figure 3.9. The facial phenotype of Hutchinson-Progeria syndrome (Figure 3.9k) is very unique and pronounced, whereas the facial phenotype of Fragile X syndrome (Figure 3.9i) is visually similar to Angelman syndrome (Figure 3.9b).

	Angelman	Apert	Cornelia de Lange	Down	Fragile X	Hutchinson-Progeria	Treacher-Collins
Apert	95.3						
Cornelia de Lange	93.1	96.2					
Down	90.5	96.6	97.4				
Fragile X	87.8	97.3	97.4	94.8			
Hutchinson-Progeria	93.5	95.5	98.5	96.7	95.3		
Treacher-Collins	94.4	95.1	97.7	98.4	98.4	97.1	
Williams-Beuren	93.6	97.9	97.3	97.8	94.2	98.1	98.4

Figure 4.4: Binary 1v1 classification accuracies (%) using late fusion of Shape SDM + FV features. On average, two syndromes can be distinguished from each other correctly 95.9% of the time.

	Angelman	Apert	Cornelia de Lange	Down	Fragile X	Hutchinson-Progeria	Treacher-Collins	Williams-Beuren
Angelman	76	1	3	4	7	0	1	8
Apert	2	94	1	1	0	0	0	2
Cornelia de Lange	2	3	91	1	0	0	0	2
Down	2	1	1	93	1	0	0	1
Fragile X	8	0	1	2	80	0	0	7
Hutchinson-Progeria	0	8	0	1	5	82	2	1
Treacher-Collins	0	2	0	0	0	0	95	4
Williams-Beuren	1	1	0	1	1	0	0	96

Figure 4.5: Normalised confusion matrix (%) for forced classification of syndromes in the LSITW test dataset using late fusion of Shape SDM + FV features. Angelman is the hardest syndrome to classify and is most frequently misclassified with Fragile X and Williams-Beuren syndromes.

4.2 Classifying Syndromes using Deep Representations

In this section we investigate the use of CNNs to perform syndrome classification on the LSITW dataset.

CNNs have been shown to reach near human-level performance on face recogni-

tion tasks. Using CNNs, in contrast to engineered-feature based representations, removes the need to handcraft features. The network can identify patterns in the training data and tune its feature representation to best represent them.

Training a CNN to perform syndrome classification from scratch would require a far greater volume of training data than we have at our disposal. We therefore make use of networks that are pre-trained on face recognition tasks as an initialisation for syndrome classifier training. This approach is called fine-tuning or transfer learning. We assume that features that are trained for facial recognition tasks are transferable to our task of recognising genetic disorders.

We contrast three CNN architectures that were pre-trained on large-scale face recognition datasets and then fine-tuned to classify syndromes (Section 4.2.1). Part-based models, where multiple models are trained for different parts of the face, were also investigated and are introduced in Section 4.2.3, as these have been shown to improve face classification tasks (Shukla et al., 2017; Gurovich et al., 2018, 2019). We present our training methodology in Section 4.2.2 and implementation details in Section 4.2.4. The results of our approach are presented in Section 4.2.5.

4.2.1 Networks and Pre-training

We compared three CNN architectures that are have shown strong performance on a variety of computer vision tasks. They were pre-trained on the VGGFace or VGGface2 datasets.

1. VGG-M pre-trained on VGGFace (Section 2.4.3)
2. ResNet50 pre-trained on VGGFace2 (Section 2.4.4)
3. Senet50 pre-trained on VGGFace2 (Section 2.4.4)

The VGG-M architecture (Chatfield et al., 2014) contains eight convolutional layers followed by three fully connected layers. The input size is 224x224x3 (width, height, channels) and the network learns a 4096-dimensional face embedding. The ResNet50 architecture is based on ResNets from (He et al., 2015) that add a residual to the output of some layers, which allows for deeper networks with more layers.

The Senet50 architecture by (Hu et al., 2017) contains additional weights that learn inter-channel dependencies. Both the ResNet50 and Senet50 are 50-layers deep, have the same input size as VGG-M and learn a 128-dimensional face embedding.

4.2.2 Fine-tuning for Syndrome Classification

The pre-trained CNNs need to be adapted to our classification task to map their feature representations onto our task of classifying 8 syndromes. Each network's face embedding layer is appended with a fully-connected layer for classification of dimensions $M \times N$, where M is equal to the dimensionality of the face embedding layer and $N = 8$ is the number of syndromes. This new classification layer is initialised randomly. This corresponds to a dimensionality of 4096x8 for VGG-M, and 128x8 for ResNet50 and Senet50 architectures. The networks were then fine-tuned to perform forced classification of the syndromes (as in Section 4.1.4).

The loss function was the softmax classification loss function shown in equation 4.8. Classes were balanced by weighting their contribution to the loss function by the inverse of the relative frequency of that class (α_k) in the training data.

$$L_{\text{softmax}}(x, y; \theta) = -\sum_{k=1}^{K} \mathbb{1}[y = k]\alpha_k \log p_k, \tag{4.8}$$

where y is the true class and p_k is the output score for sample x belonging to class k. Minimising this loss equates to maximising to the negative log score of correct classification.

The loss function is optimised using stochastic gradient descent with a constant learning rate of 0.0001 for the final classification layer and 0.00001 for all remaining layers. This accounts for the fact that the weights in the final classification layer are initialised randomly and are thus farther from the desired values than the pre-trained weights in other layers. We found that this approach worked better than freezing the weights in the pre-trained part of the network. Due to the scarcity of data, we did not use a validation set, and therefore, did not use learning rate scheduling techniques.

During training we used 50% dropout on the new classification layer of each network. Each network was trained for 45 epochs, where one epoch is a complete pass

through all images in the training dataset. Networks were tested after each epoch and the best performing checkpoint was selected for the results in Section 4.2.5.

4.2.3 Part-based Models

The methods for deep description so far considered the whole face. In this section, we investigate that use of part-based models. Facial dysmorphisms are often describe by describing abnormalities in a face part, e.g. a common phenotype of Down syndrome is "upward slanting eyes". This motivates the use of part-based models, where separate networks are trained to classify specific regions of the face. Part-based models were shown to improve face classification tasks (Shukla et al., 2017; Gurovich et al., 2018, 2019).

We trained 5 part-based models, each trained using crops around one of the following: eyes, nose, mouth, chin, and the entire face.

Face regions were defined with the help of facial landmarks that were detected using methods detailed in Section 3.2.1. The averaged location of landmarks that corresponded to a face part, e.g. the eyes, was taken as the centre point of that face part. Square crops around each part's centre point were then used as inputs to each network in the model. At inference, each crop of a test image was classified by its corresponding model and used as votes towards the final classification. This method is comparable to the late fusion experiments in Section 4.1.4.

An example of the part-based model inputs and how they are used is shown in Figure 4.6. An evaluation of these models is given in Section 4.2.5.

4.2.4 Image Augmentation

This section introduces face augmentation methods that were used when training CNNs. Image augmentation was used to artificially increase the training dataset and thus improve the learned deep feature representation of the network.

The input images were normalised by the average RGB pixel values for all images of the dataset they were pre-trained on. The dataset was augmented by applying a random combination of image transformations from the following:

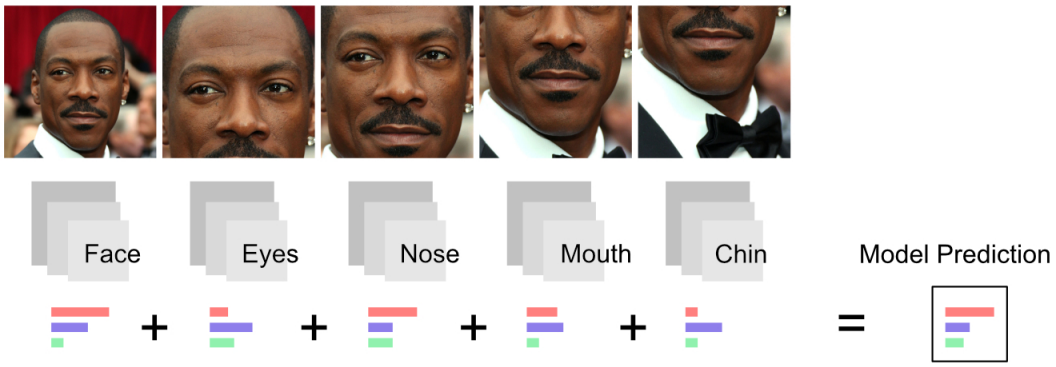


Figure 4.6: Overview of the part-based model with an example of the inputs to each network of the model: eyes, nose, mouth, chin and face. Each network's output is summed and the image is classified by the maximum across all networks.

- Crop ($p = 100\%$) - Square crop at a random location within the image of size of network input
- Rotation ($p = 100\%$) - Random rotation in the range of -10 to $+10$ degrees
- Horizontal flip ($p = 50\%$)
- Colour jitter ($p = 100\%$) - Random adjustment of brightness, contrast, hue, and saturation
- Grayscale ($p = 20\%$)
- JPEG Compression ($p = 20\%$) - JPEG compression at $95 - 100\%$ quality
- Motion blur ($p = 20\%$) - Maximum kernel size $= 7$
- Gaussian blur ($p = 20\%$) - Maximum kernel size $= 3$
- Gaussian noise ($p = 20\%$)

The probabilities p of each of these augmentations being applied is given in brackets. These augmentations account for many of the conditions we expect from images in our domain. Using these augmentations can significantly improve the generalisability of the learnt feature representation. We observed an average

Table 4.2: Summary table of average classification accuracies (percent) on test data using deep feature representations. The "PB"-prefix denotes an part-based model.

Network	Accuracy (%)
VGG-M	89.1
ResNet50	90.6
SENet50	91.8
PB-VGG-M	89.3
PB-ResNet50	91.5
PB-SENet50	**92.2**

improvement of 3% in accuracy compared to models trained without using these image augmentations.

4.2.5 Results and Discussion

Deep representations outperformed engineered-feature based representations for each architecture that was trained. The per-class averaged accuracies are shown in table 4.2. VGG-M network, pre-trained on VGGFace achieved an average classification accuracy of 89.1%, 0.7% better than the best engineered-feature based representation, namely late fusion of Shape SDM + FV (88.4% forced classification). Both ResNet50 and Senet50 architectures, pre-trained on VGGFace2 performed better, at 90.6% and 91.8%, respectively. A direct comparison between VGG-M and the other two networks is not possible, as they were pre-trained on different datasets. The lower accuracy may be due to the network being pre-trained on an "inferior" dataset; VGGFace2 is a larger dataset and contains more pose variation than VGGFace (Section 2.4.4). Alternatively, it may be due to it containing a much larger number of parameters that needed to be trained from scratch (33k vs $<$ 2.6k).

Training part-based models further improved accuracies by 0.2% (VGG-M), 0.9% (ResNet50) and 0.4% (SENet50). The best overall result was achieved when training a part-based model with the SENet50 architecture (92.2%). This indicates that there may be benefits to training networks that act as specialists for different parts of the face. The accuracies for each individual network in the part-based model are

Table 4.3: Classification accuracies of individual networks in the part-based models

Network	Face	Eyes	Nose	Mouth	Chin
E-VGG-M	89.3	85.2	83.1	77.2	61.7
E-ResNet50	91.5	88.8	88.1	85.1	65.3
E-SENet50	92.2	89.0	88.9	82.0	67.7

shown in table 4.3. The individual face parts all resulted in worse classifiers than the full face crop. On average, the top half (eyes and nose) of the face seems to be more informative for syndrome classification than the bottom half (mouth and chin).

The normalised confusion matrix for classification with SENet50 is shown in Figure 4.7 The classification accuracy improved all but one syndrome, most remarkably for Hutchinson-Progeria, which improved from 82% to 100%. One unexpected result was the drop in accuracy for classifying Williams-Beuren, 87% SENet50 vs 96% late fusion of FV+SDM. They were, however, confused with Angelman syndrome and Fragile X syndrome, which proved difficult to classify with engineered-feature based representations as well.

	Angelman	Apert	Cornelia de Lange	Down	Fragile X	Hutchinson-Progeria	Treacher-Collins	Williams-Beuren
Angelman	78	0	0	0	7	0	3	12
Apert	0	96	0	2	0	0	2	0
Cornelia de Lange	2	2	96	0	0	0	0	0
Down	0	0	0	97	3	0	0	0
Fragile X	6	0	0	6	85	0	0	3
Hutchinson-Progeria	0	0	0	0	0	100	0	0
Treacher-Collins	0	5	0	0	0	0	95	0
Williams-Beuren	9	0	0	0	4	0	0	87

Figure 4.7: Normalised confusion matrix for forced classification of syndromes in the LSITW test dataset using SENet50 pre-trained on VGGFace2. The color-scale is the same as that in Figure 4.5. Angelman is the most difficult syndrome to classify and Hutchinson-Progeria is correctly classified with 100% accuracy.

Average classification accuracies were used in this chapter to contrast the relative performance of each method. Although they are a good high-level indicator of

relative performance, they do not provide detailed explanations for the aggregate score. The confusion matrices in this chapter go some way toward understanding which syndromes were most often confused. Furthermore, gradient-based attention maps (Selvaraju et al., 2017) can be used to visualise the parts of the image that were most informative for a CNN's decision. If these regions coincide with the known facial phenotype, that would confirm that these are indeed what the network is basing its decision on.

4.3 Conclusions

This chapter discussed classification of genetic disorders with engineered-feature based and deep representations and compared these to previous work by Ferry et al. (2014) (Section 4.1.1).

We showed that shape descriptors using SDM were a significant improvement upon previous work, as the predicted locations of the landmarks were more accurate. Dense description of the face using Fisher Vector encoded SIFT descriptors, proved to be the best of the engineered-feature based descriptors. Late data fusion, by summing the outputs of classifiers trained separately on Shape SDM and FV resulted in the highest classification accuracies for engineered-feature based representations.

Neural networks, when pre-trained on large face recognition datasets and fine-tuned for syndrome classification, were shown to outperform all engineered-feature based methods . This shows that deep representations for face recognition can be appropriated and adapted for syndrome classification. Furthermore, part-based models where each network specialises in classifying different parts of the face further improve classification results.

Angelman syndrome and FragileX syndrome proved the most difficult to classify in all experiments, despite a relatively large number of training images for Angelman syndrome. We hypothesise this to be due to the facial phenotypes of both syndromes being less prominent than that of the other syndromes that were studied in this chapter. The opposite effect was witnessed when classifying the more visually

distinct phenotype of Hutchinson-Progeria, which was classified perfectly in the test set using the SENet50 network.

5

Removing Biases from Deep Neural Network Embeddings

Contents

5.1 Related Work 82
5.2 Datasets 84
5.2.1 LAOFIW: Labeled Ancestral Origin Faces in the Wild . 84
5.2.2 "Cleaned" IMDB Dataset 86
5.2.3 Annotated Faces in the Wild 88
5.3 Unbiased Learning Algorithm 88
5.4 Experiments on Removing Bias 91
5.4.1 Removal of a bias from a network 92
5.4.2 Removal of an extreme bias from a network 92
5.4.3 Simultaneous removal of multiple spurious variations . . 93
5.4.4 Implementation details 93
5.5 Results 94
5.5.1 Removal of a bias from a dataset 94
5.5.2 Removal of an extreme bias from a network 96
5.5.3 Simultaneous removal of multiple spurious variations . . 97
5.6 Conclusion 101

In this chapter, we present work on bias removal from the feature representation of convolutional neural networks.

CNNs are usually trained on large datasets, that contain a number of biases or spurious variations (Tommasi et al., 2015; Torralba and Efros, 2011). CNNs

have been shown to "cheat", by learning, and in some cases amplifying, biases present in training data (Zhao et al., 2017; Burns et al., 2018) , e.g. training on a dataset depicting women in kitchens more often than men would result in a network learning that being in the kitchen was a key feature for the identification of women. Though this may have been true for the dataset the classifier was trained on, in general, this is not a reliable indicator of the presence of a woman.

These biases also exist in face recognition datasets, which are often composed of celebrities (Cao et al., 2018; Guo et al., 2016; Huang et al., 2007; Kemelmacher et al., 2015; Kostinger et al., 2011; Parkhi et al., 2015). These can contain age, gender, and ancestral origin biases, e.g. female celebrities tend to be younger than their male counterparts. This bias does not represent the real world outside of the movie business, reducing the usefulness of models trained on these datasets. Furthermore, networks were shown to learn biases, even when datasets were balanced purposely (Wang et al., 2018b).

Furthermore, when conducting transfer learning we can unknowingly inherit those biases and carry them forward to a new task. This is particularly relevant to the domain of genetic disorders, where, due to the scarcity of training data, transfer learning is an important pre-requisite for using deep learning. Furthermore, there may exist known biases, e.g. all examplars of a disorder may be photographed under the same conditions at a single clinical facility. A network could resort to "cheating" and e.g. memorise a shared background for a group of images. This would render the network useless for use in real-world clinical applications, unless this bias could be removed.

Our bias removal algorithm is inspired by a domain and task adaptation approach (Tzeng et al., 2015), that can be used to 1) to ensure that a network is blind to a known bias in the dataset, 2) improve the classification performance when faced with an extreme bias, and 3) remove multiple spurious variations from the feature representation of a primary classification task. We use age, gender, ancestral origin, and pose information for facial images to demonstrate our framework.

Biases towards certain ancestral origins have been shown to affect face detection models, with darker skinned females most often misclassified (Buolamwini and Gebru, 2018). In order to investigate the effect of racial biases on the feature representation of CNNs, we created a new labelled ancestral origin dataset, "Labelled Ancestral Origin Faces in the Wild (LAOFIW)", from publicly available images. This dataset contains 14,000 facial images of individuals whose ancestral origin is one of sub-Saharan Africa, Indian Subcontinent, Europe, and East Asia, with a variety of ages, poses, illumination, and facial expressions.

The rest of this chapter is organized as follows. Section 5.2 introduces our new LAOFIW dataset and other datasets used as part of this work. Section 5.3 discusses the methods to remove spurious variations from the feature representation of a network. In Section 5.4.1, we present an experiment to remove a bias from the feature representation network. In Section 5.4.2, we investigate how removing an extreme bias can improve classification performance, and in Section 5.4.3, we investigate removing multiple spurious variations from a network. The results are detailed in Section 5.5 and, finally, summarised in Section 5.6.

5.1 Related Work

Image datasets are known to contain biases that cause models to generalize poorly to new, unseen data (Torralba and Efros, 2011; Tommasi et al., 2015). This has been addressed by domain adaptation that aims to minimize the difference between the source and target domains (Gretton et al., 2009). Domain adaptation has been shown to improve the generalisability of classifiers for the same task across different domains (Tzeng et al., 2015; Tzeng et al., 2017). We draw inspiration from domain adaptation methods to make CNNs agnostic to spurious variations.

Learning of feature representations that are invariant to certain spurious variations has been tackled in a number of ways. Yin and Liu (2018) take a multi-task learning approach to pose-invariant face recognition. They propose a novel approach to automatically weight each of the spurious variations during training. Another approach is to adjust the training data distribution at training time, to avoid

learning biases (Zhao et al., 2017). This method relies on having labels for each of the spurious variations for each training instance. This is not feasible for most existing datasets, as they tend to be labeled with a single task in mind, where information about spurious variations is not available. Our approach can make use of separate datasets, each labelled for distinct tasks, to remove multiple spurious variations simultaneously.

Slavin Ross et al. (2017) and Raff and Sylvester (2018) draw inspiration from Ganin et al. (2017) to remove a source of variation with the use of a gradient reversal layer to update the network in opposition of a task. Instead of applying gradient reversal on the output of the softmax layer, which penalizes correct classifications, we compute the cross-entropy of the output classifier and a uniform distribution, as in Tzeng et al. (2015); Tzeng et al. (2017). This ensures an equally uninformative classifier across all tasks.

Further research was conducted on bias removal since this work was published. These are summarised below.

Prabhu et al. (2019) investigated the effect of inter-attribute biases within feature representations of networks using the CelebA dataset (Liu et al., 2014) that contains 40 semantic facial attributes. When performing face manipulation by interpolating along the latent vector corresponding to more "attractive" faces, input faces, regardless of ancestral origin or sex were transformed to young, blond females. This was undone in a post-hoc manner by modelling the relationships between attributes as a graph and applying Markov-blanket correction factors to the latent space vectors.

An approach to reduce dataset bias by introducing synthetic data in pre-training was investigated by Kortylewski et al. (2019). They use a face image generator to generate faces for known nuisance variables and pre-trained networks on highly variable image. This was followed by fine-tuning on real-world data. The resulting networks were shown to be more generalisable.

A new face dataset for ancestral origin classification was collected by Wang et al. (2018a) consisting of about 40,000 images across four broad ancestral origins. The

images were curated by selecting images from within the MS-Celeb-1M dataset, which contains 10 million faces of 1 million public figures (Guo et al., 2016).

5.2 Datasets

5.2.1 LAOFIW: Labeled Ancestral Origin Faces in the Wild

A new ancestral origin database was created as part of this work called "Labeled Ancestral Origin Faces in the Wild (LAOFIW)". The aim of this dataset was to 1) create a dataset for experimentation, and 2) be used as a spurious variation dataset for applications where training a network to be agnostic to ancestral origin is important.

The database was assembled using the Bing Image Search API. Search terms were selected by geographical regions, e.g. for the Western European class "German, English, Polish, etc". These terms were submitted to the search engine in conjunction with the words "man, woman, boy, girl". Additionally, results were filtered using API parameters to return only photographic images of faces. A complete list of the search terms we used is given in table 5.1. In total, 42 origin search terms were queried returning 20,000 images. A significant number of these images were duplicates. As duplicates often varied slightly due to differences in compression and crops, simple pixel-to-pixel comparisons between images could not be used to identify them. Instead, we computed Histogram of Oriented Gradients features (Dalal and Triggs, 2005) for each image and encoded them using a Gaussian mixture model with 512 centres to create an image "signature". Images with the same signature were classed as duplicates and removed. The remaining 14,000 images, were manually divided into four broad ancestral origins: sub-Saharan Africa, Indian Subcontinent, Europe, and East Asia. These classes were selected on the basis of being visually distinct from each other.

Images were selected so that male and female individuals were represented in similar numbers. The data is diverse in backgrounds, lighting and expression, which is reflected in our coice to name it "...in the Wild". It also contains images with

Table 5.1: Table of search terms for LAOFIW

German	Chinese	African	Pakistani
Russian	Tibetan	Sudanese	Indian
Italian	Korean	Congolese	Bengali
French	Taiwanese	Kenyan	Mumbai
English	Hong Kong	Nigerian	Lahore
Irish	Mongolian	Zambian	Delhi
Spanish	Japanese	Zimbabwean	Kerala
Austrian	Beijing	Ugandan	Gujarat
Swiss	Shanghai	Eritrean	Marathi
Polish	Guangzhou	Ivory Coast	Kashmiri

varied poses: more than a third of the images are non-frontal. Sample images are shown in Figure 5.1. This dataset is available upon request.



Figure 5.1: Example images from the LAOFIW dataset for each of the four classes. In rows top to bottom: Sub-Saharan Africa, Europe, East Asia, Indian Subcontinent. The images are highly variable in age, gender, pose, lighting and expression.

The LAOFIW database was used to investigate the effect of racial bias in the feature representation of networks. Of the 14,000 images, 12,000 were were used

for training (4,000 for each ancestral origin) and the remaining 2,000 images (also class-balanced) were retained as a test dataset.

5.2.2 "Cleaned" IMDB Dataset

The IMDB dataset is a large publicly available dataset of the images contained in the profiles of the 100,000 most popular actors on the IMDb website[1] with gender and date of birth labels provided for each instance (Rothe et al., 2015). This dataset was used to investigate the effects of age and gender bias in celebrity datasets.

The labels in the IMDB dataset are noisy, with a number of images having both incorrect age and gender labels. This is due to the nature of how the data was collected: the authors assumed that images from an actor's profile that contained a single face would contain the actor in question. As stated by the authors, however, these images often contain other actors, who co-starred in their movies (Rothe et al., 2015). As gender and date of birth are taken from the profile of the actor, this causes erroneous labels for images that show co-stars. The age of an individual was calculated as the difference between the timestamp of the photo and the date of birth of the subject. In some cases, the time stamps of photos pre-date an actor's date of birth or are otherwise unreliable.

To mitigate this problem, we used the Microsoft Azure Face API[2] to extract gender and age estimates for an identity-balanced subset of 150,000 images. All images in which the predicted gender from Azure and the IMDB gender label disagreed were rejected. We then ran the same analysis for age and removed images in which the Azure age prediction differed by more than 10 years from the IMDB labels. The resulting cleaned dataset contained 60,000 images.

In order to quantify the effectiveness of this data cleansing procedure, we trained a gender classification network on both the original 150,000 images and the cleaned 60,000 images, withholding 20% of the images for testing. We made use of a VGG-M network (Chatfield et al., 2014), pre-trained on the VGGFace dataset

[1] `https://www.imdb.com/`

[2] `https://azure.microsoft.com/en-gb/services/cognitive-services/face/`

(Section 2.4.3). Gender classification accuracies on the test images improved from 75%, before cleaning, to 99% after cleaning.

We made used the cleaned IMDB database to investigate age and gender biases. The distribution of ages for each gender in the cleaned IMDB dataset is shown in Figure 5.2. It can observed that a bias towards younger women and older men is present in the data. This reflects reality in the film industry, where women generally have shorter careers than their male counterparts.

A subset of the cleaned data was used to create an unbiased test dataset, which contains equal numbers of men and women for each age category. This unbiased test dataset was used to evaluate network bias and was not used during training.

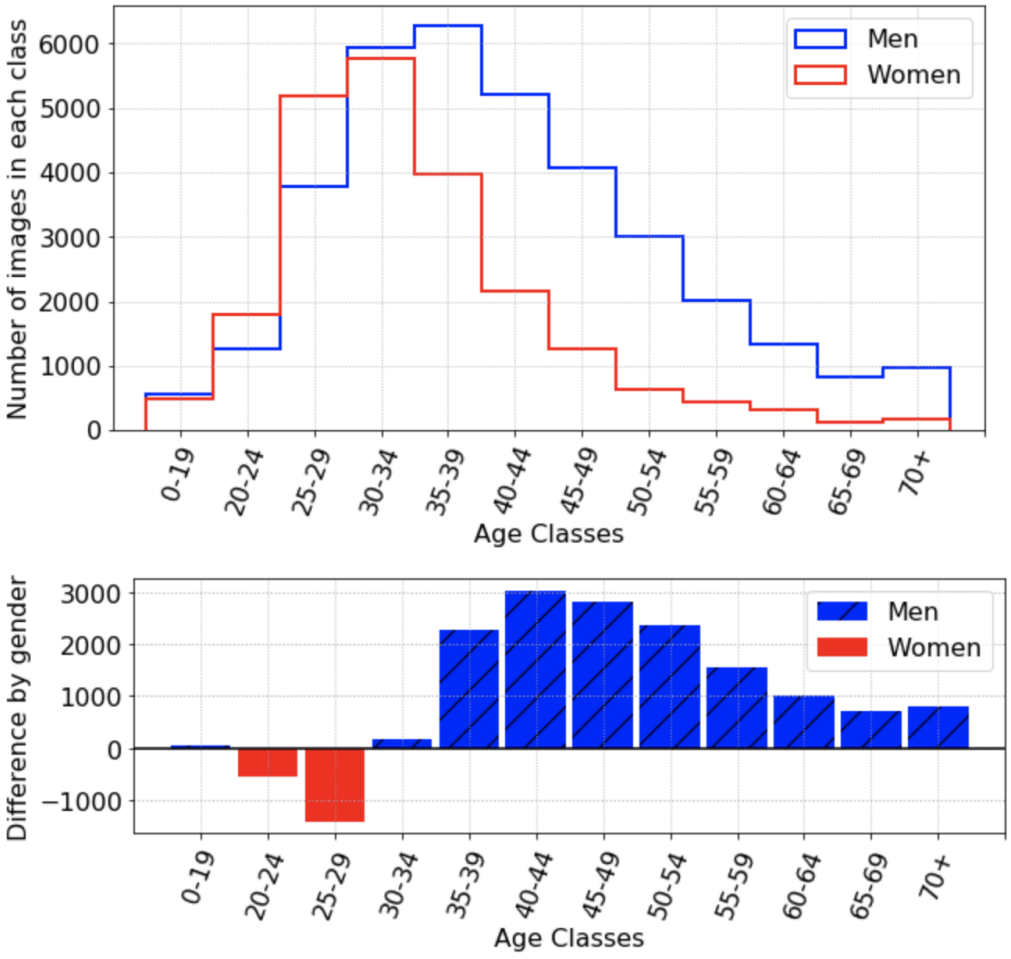


Figure 5.2: Age distributions and differences for women (red) and men (blue) in the cleaned IMDB dataset. The dataset bias towards younger women and older men is visible.

5.2.3 Annotated Faces in the Wild

Annotated Faces in the Wild (AFLW) is a face dataset that contains 25,993 images of faces collected from Flickr (Kostinger et al., 2011). These images contain 21 manually annotated facial landmarks and are categorized by sex. The dataset also contains estimated face pose angles (pitch, yaw, and roll from neutral).

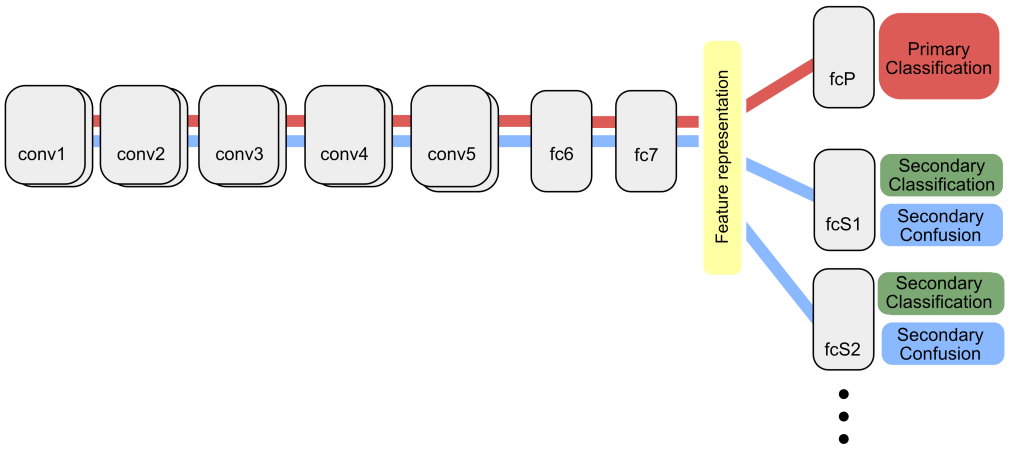
We made use of the AFLW dataset categorized by yaw, to classify the pose of faces. The poses were binned into five categories, "profile left, near-frontal left, frontal, near-frontal right, profile right". The non-frontal images were duplicated and flipped to augment the dataset. A class-balanced subset of 24,000 images was selected for training, and a class-balanced set of 6,000 images was reserved for testing.

5.3 Unbiased Learning Algorithm

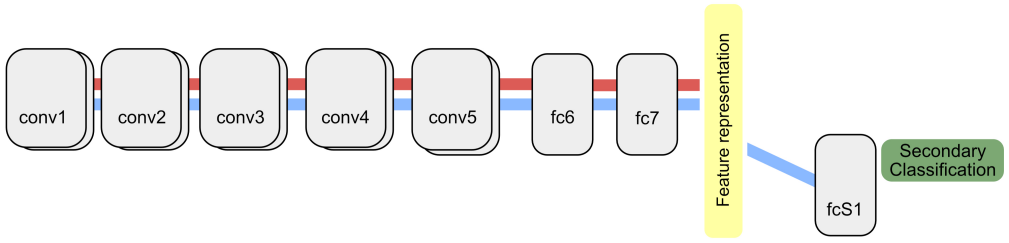
In the following, we introduce our joint learning and unlearning (JLU) algorithm that aims to learn a single feature representation θ_{repr}, that is informative for a primary classification task, whilst simultaneously being uninformative for a number of spurious variations, that represent undesirable sources of variation. For example, we may wish to create an age classifier, that does not base its decisions on any pose, ancestral origin, or gender information.

The convolutional network (CNN) architecture is depicted in Figure 5.3. The primary branch has a single classification loss, the primary classification loss, which assesses the ability of the network to accurately distinguish between classes of the primary task. Each secondary branch has two losses: a classification loss and a confusion loss. These losses are used to, in turn, assess the amount of spurious variation information left in the feature representation θ_{repr} and then remove it.

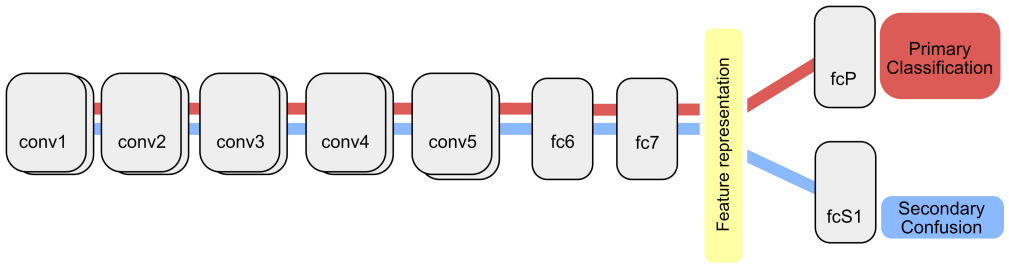
Assume access to a primary dataset $\mathcal{D}_p = \{x_i, y_i\}_{i=1}^{n_p}$, containing n_p images, x_i, with labels $y_i \in \{1, ..., K\}$. And similarly, assume access to M secondary datasets, $\mathcal{D}_s = \{\mathcal{D}_m\}_{m=1}^{M}$, each describing a single spurious variation.



(a) Overview of CNN Architecture



(b) Secondary classification



(c) Primary classification and secondary confusion

Figure 5.3: We use a joint loss over primary and secondary data to learn a feature representation that simultaneously learns to classify the primary task but becomes invariant to secondary tasks, the spurious variations. The spurious variation (secondary) classification loss and confusion loss act in opposition to learn the classifier on the feature embedding and change the feature embedding to confuse the classifier, respectively. They are optimised alternatively according to Algorithm 1, as shown in (b) and (c). The base architecture is a VGG-M network (Chatfield et al., 2014).

Let the classification objective for a generic task with K classes and a corresponding classifier θ_c, given a feature representation θ_{repr}, be defined as the standard softmax loss:

$$L_{\text{softmax}}(x, y; \theta_{\text{repr}}, \theta_c) = -\sum_{k=1}^{K} \mathbb{1}[y = k] \log p_k, \tag{5.1}$$

where p_k is the softmax of the classifier's output assigning the input x to class k. We will refer to this loss evaluated on the primary task as $L_P(x_p, y_p; \theta_{\text{repr}}, \theta_P)$ and on the m-th spurious variation as $L_m(x_m, y_m, \theta_{\text{repr}}; \theta_m)$.

Inspired by Tzeng et al. (2015), we introduce a confusion loss for the m-th spurious variation, $L_{\text{conf}}(x_m, y_m, \theta_m; \theta_{\text{repr}})$, in (5.2). Minimizing the confusion loss seeks to change the feature representation θ_{repr}, such that it becomes invariant to the spurious variations.

$$L_{\text{conf,m}}(x_m, y_m, \theta_m; \theta_{\text{repr}}) = -\sum_{m=1} \frac{1}{m} \log p_m. \tag{5.2}$$

We compute the best classifier for each spurious variation, θ_m, and then compute the entropy between the output predicted from each of these classifiers and a uniform distribution. The complete method minimizes the joint loss function:

$$\begin{aligned} L(x_p, y_p, x_s, y_s, \theta_p, \theta_s, \theta_{\text{repr}}) = L_p(x_p, y_p; \theta_{\text{repr}}, \theta_p) \\ + L_s + \alpha L_{\text{conf}}, \end{aligned} \tag{5.3}$$

where $0 < \alpha < 1$ and determines how strongly the confusion loss affects the overall loss, $\theta_s = \{\theta_1, ..., \theta_M\}$, and

$$L_{\text{conf}} = \frac{1}{M} \sum_{m=1}^{M} L_{\text{conf},m}(x_m, y_m, \theta_m; \theta_{\text{repr}}), \tag{5.4}$$

$$L_s = \sum_{m=1}^{M} \beta_m L_m(x_m, y_m, \theta_{\text{repr}}; \theta_m), \tag{5.5}$$

where β_m is a weight assigned to the $m-$th spurious variation. As mentioned in (Tzeng et al., 2015), the confusion loss (5.4) and the spurious variation classification loss (5.5) stand in opposition to one another, so they cannot be optimized in the same step. Therefore, we switch between training the spurious variation classification loss, L_s, and then the joint primary and confusion loss, $L_p(x_p, y_p; \theta_{\text{repr}}, \theta_p) + \alpha L_{conf}$. At each iteration, the best spurious variation classifier is found for the feature representation. This spurious variation classifier tells us how much information

thereof remains in the feature representation of the network. We then attempt to remove it with the confusion loss. The training procedure is shown in algorithm 1.

We used the VGG-M architecture (Chatfield et al., 2014), which consists of five convolutional layers (conv1-5) and three fully connected layers (fc6-8). The feature representation parameters θ_{repr} represent layers conv1-fc7.

Algorithm 1 Joint Learning and Unlearning

procedure MINIMISE $\mathrm{L}(x_p, y_p, x_s, y_s, \theta_p, \theta_s, \theta_{\text{repr}})$
  for epochs **do**
    while $\frac{dL_s}{d\theta_s} \neq 0$ **do**
      minimise L_s
    end while
    minimise $L_p + \alpha L_{\text{conf}}$
  end for
end procedure

5.4 Experiments on Removing Bias

In this section, we present experiments to demonstrate possible applications of our methodology and the datasets they were tested on.

Removal of a bias from a network — Training a gender-agnostic age classifier using the IMDB dataset, to overcome a gender bias: female celebrities tend to be younger than their male counterparts in this dataset.

Removal of an extreme bias from a network — Training an age-agnostic gender classifier on subsets of the IMDB dataset that contain only young women and only old men, and vice versa.

Simultaneous removal of multiple spurious variations — Demonstrating our JLU algorithm's ability to simultaneously remove multiple spurious variations from the feature representation of a network trained for a primary classification task.

5.4.1 Removal of a bias from a network

We investigated the task of creating an age classifier using the cleaned IMDB dataset described in Section 5.2.2. We hypothesize that the gender bias in the distribution, where men are generally older than their female counterparts in this dataset, will be learned by the network. Since the training dataset is relatively small for a deep learning task, age estimation is not approached as a regression task. Instead, age classification is conducted by creating age bins of 5 years, or more years at either end of the distribution. The categories are shown in Figure 5.2. Predictions within one class from the true age class are accepted as positive classifications to account for errors caused by edge cases in different bins.

Two VGG-M networks were trained to perform this task: 1) baseline — trained solely on age data, 2) blind — using the JLU algorithm to trained on age data whilst removing gender-specific information from the network.

Both networks were evaluated on the unbiased test dataset detailed in Section 5.2.2. Accuracies for age classification and prediction distributions were compared for both genders.

5.4.2 Removal of an extreme bias from a network

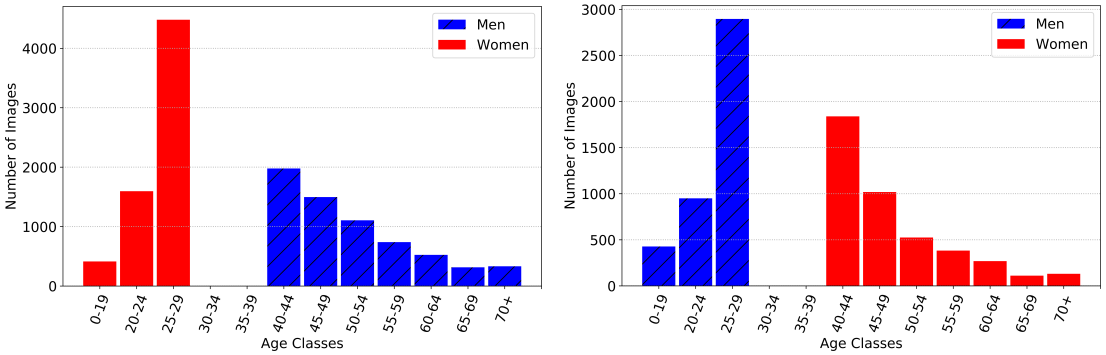
Gender classifiers were trained using the following artificially biased subsets of the cleaned IMDB data from Section 5.2.2:

1. Extreme bias 1 (EB1): women aged 0-30, men aged 40+

2. Extreme bias 2 (EB2): women aged 40+, men aged 0-30

A buffer of 10 years without any individuals was placed between the two sets to exaggerate the age difference between the genders. We hypothesize that the age bias in the distribution will be learned by the network, for example, if trained on EB1, it should falsely learn to predict that young men are in fact women. Histograms of these datasets are given in Figure 5.4.

We trained a baseline network on gender only, and a blind network to unlearn age with our JLU algorithm. The trained classifiers were evaluated on the unbiased test dataset detailed in Section 5.2.2.



(a) EB1: women aged 0-30, men aged 40+ **(b)** EB2: women aged 40+, men aged 0-30

Figure 5.4: Histograms showing the ages of women (red) and men (blue) in the artificially biased datasets. In both cases, individuals aged under 35 were selected for the young, and individuals aged above 45 were selected for the old set.

5.4.3 Simultaneous removal of multiple spurious variations

In this experiment we will investigate the efficacy of our JLU algorithm for unlearning multiple spurious variations from the feature representation of a network. The primary task and spurious variations are selected from age, gender, ancestral origin and pose. In each case, one task is selected as the primary task, and the remaining three make up the spurious variations that are to be removed. As before, a baseline network was trained for each primary task without using our JLU algorithm.

The cleaned IMDB dataset described in Section 5.2.2 was used for age and gender labels. The LAOFIW dataset, described in Section 5.2.1 was used to classify ancestral origin. Finally, the adapted AFLW dataset (Section 5.2.3) was used to classify pose.

5.4.4 Implementation details

The base network is an adapted VGG-M network (Chatfield et al., 2014), pre-trained on the VGG-Face dataset 2.4.3. In our experiments, we noticed no significant improvement in updating the weights in the convolutional layers, therefore, the

weights in layers conv1-conv5 were frozen for all experiments. This significantly increased the speed of the training algorithm.

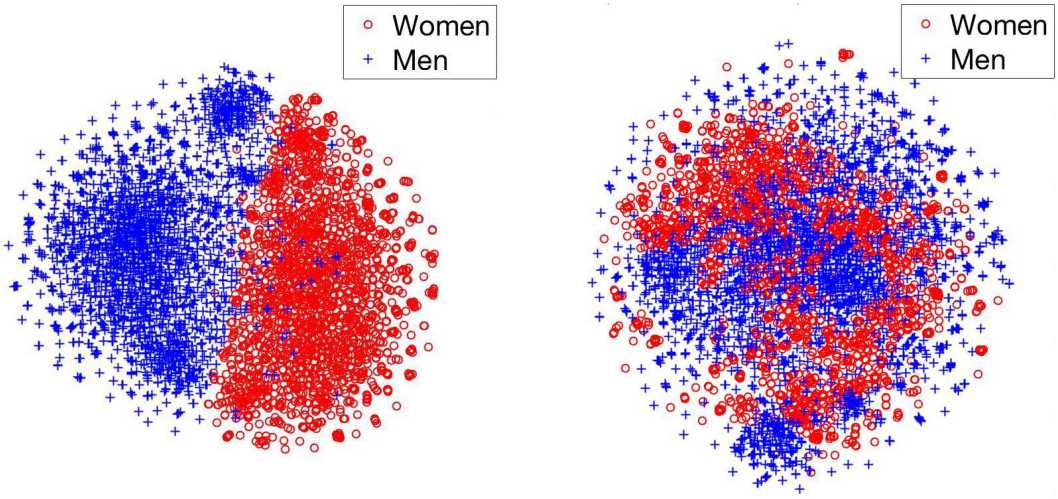
The confusion loss was approximated using the Kullback-Leibler divergence between the softmax output of a classifier and a uniform distribution. The network was trained using stochastic gradient descent (SGD) with a constant learning rate of 1×10^{-4}. Learning rates of classification layers for each task were boosted by a factor of 10. The hyperparameter in Equation 5.3 was found using a grid search and set to a value of $\alpha = 0.1$. The spurious variation classifier hyperparameters were all set to $\beta_m = 1$ to weight them equally. To address the imbalance in the class distributions of the training data, losses were weighted by the inverse of the relative frequency of that class. The MatConvNet framework (Vedaldi and Lenc, 2015) was used for these experiments.

5.5 Results

5.5.1 Removal of a bias from a dataset

We computed feature representations for the class-balanced test dataset from Section 5.2.2 for both baseline and blind networks. T-SNE visualizations of these feature embeddings are shown in Figure 5.5. The feature representation of the baseline network that was trained to classify age is clearly separable by gender, demonstrating that the bias in the training data was learned. After unlearning gender, the feature representation is no longer separable by gender, demonstrating that this bias has been removed.

The distributions of the age predictions and ground truth on the gender-balanced test set are shown in Figure 5.6. The red and blue lines show the predicted ages for women and men, respectively. The black dashed line shows the ground truth for both genders. The distributions of predicted ages for females and males are different for the baseline network (Figure 5.6a): women's ages tend to be underestimated, whereas men's ages are generally overestimated. This behaviour mimics the bias within the training data, which confirms that the classification model is making age predictions that are, to a degree, dependent on gender.

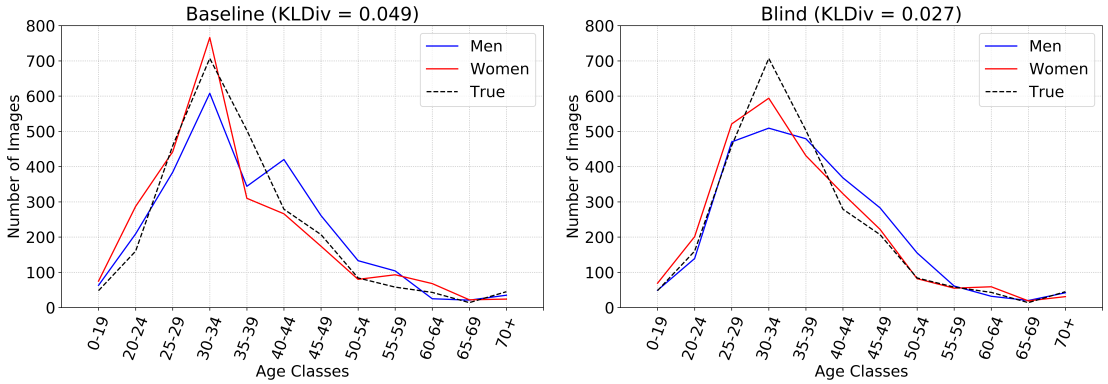


(a) Feature representation of baseline network **(b)** Feature representation of blind network

Figure 5.5: T-SNE visualizations of 4096-dimensional feature representation of the class-balanced test dataset for age classification networks trained on gender-biased data. (a) Baseline network without unlearning gender, (b) Blind network with gender unlearning. The feature representation is clearly separable by gender for the baseline network, showing that the network has learned the gender bias in the dataset. After unlearning, this bias is no longer pronounced.

In order to be sure that the network is not using gender information to predict age, both female and male prediction distributions should be similar. Figure 5.6b shows the distributions for the network that has been trained to no longer be able to differentiate between genders. The KL-divergence between age prediction distributions for men and women of the biased and unbiased networks are 0.049 and 0.027, respectively. The reduction in KL-divergence shows that the network has successfully unlearned gender, as the predicted age distributions for both genders are similar. Note, that we are not necessarily trying to perfectly predict the ground truth, but be confident that men and women are not being treated differently.

The average prediction accuracy for age classification on the gender-balanced test dataset for the baseline network was 78.9% (78.4% for females and 79.4% for males) and for the unbiased network was 78.1% (77.4% for females and 78.9% for males). This corresponds to a reduction in accuracy of 0.8% (1.2% for females and 0.5% for males).



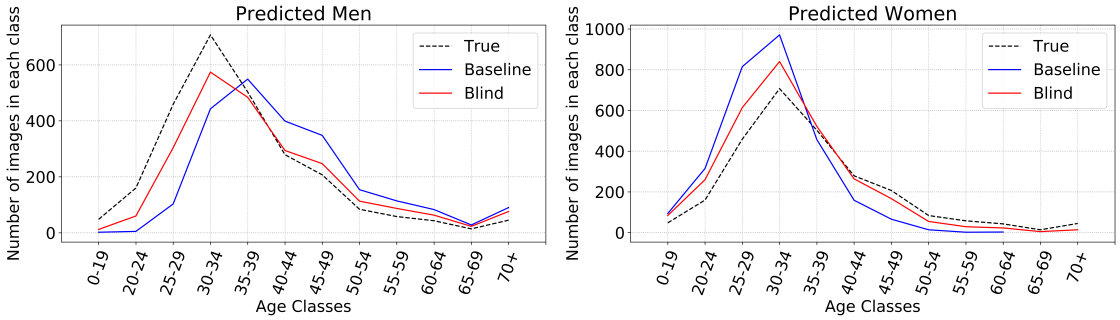
(a) Baseline network **(b)** Unbiased network

Figure 5.6: Prediction distributions of age for each gender compared to the ground truth (black) for men (blue) and women (red). The evaluation dataset was gender-balanced so that each class has an equal number of women and men. (a) Baseline network without unlearning gender, (b) unbiased network with gender unlearning. The prediction distributions for women and men align after unlearning gender. The KL-divergence score between the prediction distributions of men and women are given for each case. The KL-Divergence score reduces, as the two distributions become more similar.

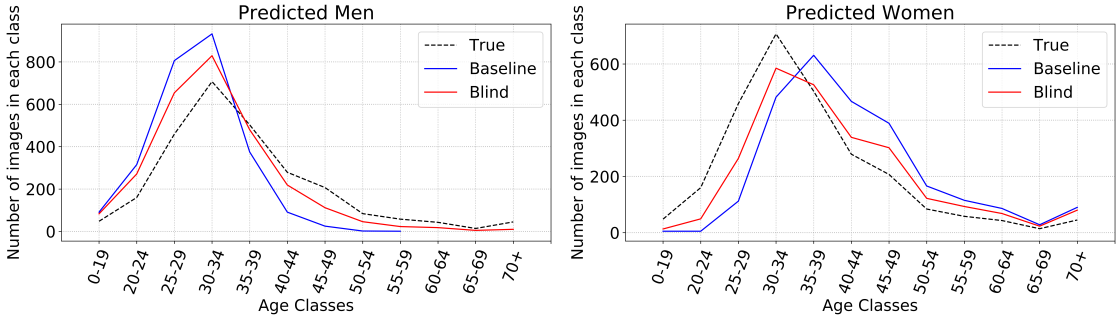
5.5.2 Removal of an extreme bias from a network

The baseline gender classification accuracy on the gender-balanced test data was 70% for a network trained on the EB1 dataset. This was improved by 16% to 86% for the blind network, where age information was simultaneously unlearned. The accuracies for networks trained on the EB2 dataset were 62% for the baseline and 82% for the blind network. This amounts to a 20% increase in classification accuracy.

The age distributions of individuals that were predicted to be either male or female, trained on the EB1 dataset, are shown in Figure 5.7. The baseline network often wrongly predicts older individuals to be male, and younger individuals to be female, in line with the training data. The age distribution of gender predictions from the JLU network, however, is closer to the true distribution. We observed similar results for the baseline network trained on the EB2 dataset, where younger individuals were more often predicted to be male, and older individuals were more often predicted to be female. The JLU network prediction distributions were closer to the true distributions.



(a) Male and female predictions by age by network trained on EB1 dataset



(b) Male and female predictions by age by network trained on EB2 dataset

Figure 5.7: Age distributions of male and female predictions from networks trained on the biased (a) EB1 and (b) EB2 datasets, evaluated on the gender-balanced test data. The baseline network (blue) often wrongly predicts older individuals to be male and younger individuals to be female. Using the JLU algorithm (red) reduces this bias and shifts the prediction distributions closer to the true distribution (black dotted).

5.5.3 Simultaneous removal of multiple spurious variations

Figure 5.8 shows how the accuracies on the test data for each attribute vary over the JLU training procedure when simultaneously unlearning multiple spurious variations. For clarity, in the figure, spurious variation classification accuracies were rescaled using Equation 5.6.

$$a = 1 - \frac{e}{e_{\max}}, \tag{5.6}$$

where e is the mean-class error of the classifier and $e_{\max}$ corresponds to the error-rate of a classifier that draws at random. A perfect classifier corresponds to a score of $a = 1$ and a random classifier corresponds to a score of $a = 0$. The primary classification accuracy was not rescaled.

These results are summarised in table 5.2 for networks trained with and without JLU. The baseline column corresponds to the accuracy of the best classifier on a feature embedding trained without JLU. The "blind" column shows the same accuracies when using the JLU algorithm. When the classifier cannot recover meaningful information from the feature embedding the accuracy is equivalent to random chance.

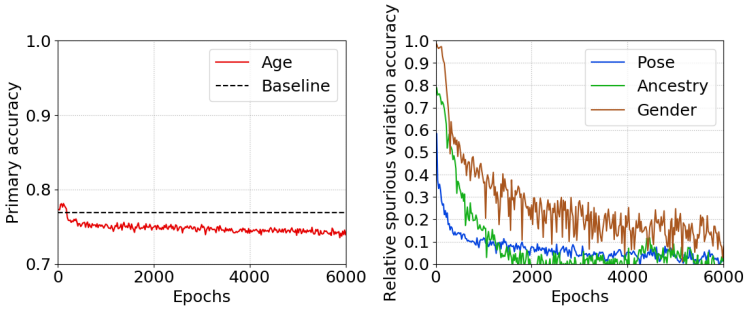
Age — The primary classification accuracy of the blind network is 5% less than the baseline network. Information about ancestral origin was removed from the network completely. Gender and pose information were removed by 88% and 91%, respectively.

Ancestral Origin — The primary classification accuracy of the blind network improved by 1% compared to the baseline network. The proportions of age, gender, and pose information that were unlearned were 92%, 98%, and 86%, respectively.

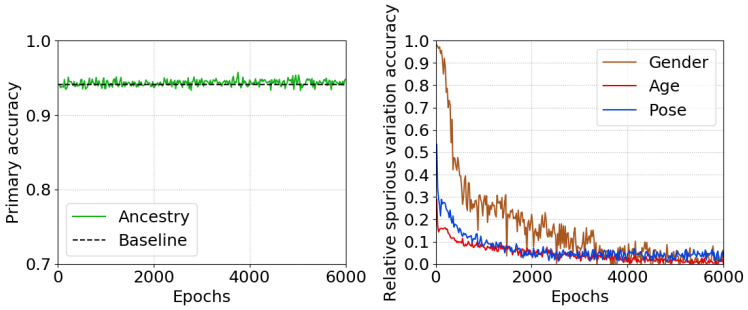
Gender — The primary classification accuracy of the blind network is the same as the baseline network. Age, ancestral origin, and pose information were removed by 94%, 98% and 95%, respectively.

Pose — The primary classification accuracy of the blind network is the same as the baseline network. Information about age, ancestral origin, and gender were removed from the network by 93%, 78% and 86%, respectively.

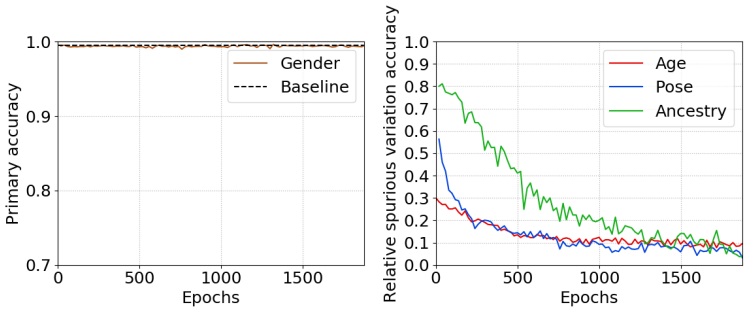
Apart from the network that was trained on the primary task of age classification on data that contained a strong gender bias, all blind networks have the same or better primary classification accuracy compared to their baseline counterparts. Spurious variations were successfully removed in most cases, with classification accuracies reducing to within 5% of random chance in 9/12 cases. The pose experiment proved the least effective, with ancestral origin information proving most difficult to remove.



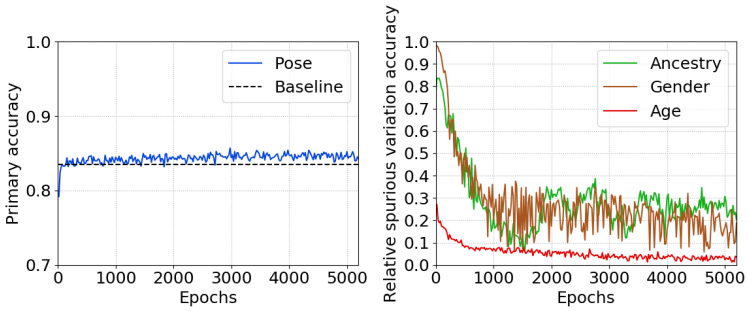
(a) Age — Ancestral Origin, Gender, Pose



(b) Ancestral Origin — Age, Gender, Pose



(c) Gender — Age, Ancestral Origin, Pose



(d) Pose — Age, Ancestral Origin, Gender

Figure 5.8: Classification accuracies on test data for the primary task and spurious variations during the training procedure of the JLU algorithm. The primary classification accuracy is compared to the baseline network accuracy (dotted). The spurious variation accuracies have been re-scaled using Equation 5.6 so that zero accuracy corresponds to the accuracy of a random chance classifier.

Table 5.2: Classification accuracies for each spurious variation for a baseline network and a "blind network". The baseline accuracies show the best classifiers that can be learned on the feature representation of the primary task without the JLU algorithm. The "Blind" column shows the classification accuracies after JLU. The random chance column states the target accuracy for a spurious variation classifier after JLU, corresponding to an uninformative classifier's accuracy. The % unlearned column shows the relative difference of baseline and blind accuracies to random chance.

Primary	Spurious	Baseline	Blind	% Unlearned	Random Chance
Age		0.79	0.74		
	Ancestral Origin	0.83	0.25	100%	0.25
	Gender	0.99	0.56	88%	0.50
	Pose	0.66	0.24	91%	0.20
Ancestral Origin		0.94	0.95		
	Age	0.58	0.12	92%	0.08
	Gender	0.99	0.51	98%	0.50
	Pose	0.56	0.25	86%	0.20
Gender		0.99	0.99		
	Age	0.70	0.12	94%	0.08
	Ancestral Origin	0.85	0.26	98%	0.25
	Pose	0.62	0.22	95%	0.20
Pose		0.84	0.84		
	Age	0.68	0.12	93%	0.08
	Ancestral Origin	0.88	0.39	78%	0.25
	Gender	0.99	0.57	86%	0.50

5.6 Conclusion

This chapter presents an approach for removing multiple known dataset biases or spurious variations from a deep neural network. Similar to previous work in the field, the algorithm is inspired by domain adaptation work, such as Tzeng et al. (2015). We compare the cross-entropy between the output distribution of classifiers that are trained to predict spurious variations and a uniform distribution. The resulting feature representation is informative for the primary task but blind to one or more spurious variations.

We demonstrated our algorithm's efficacy on face classification tasks of age, ancestral origin, pose and gender. The resulting feature representations remained informative for one task, whilst simultaneously being uninformative for the others. When training a gender classification network on extremely age-biased data, our algorithm significantly improves (by up to 20%) classification accuracies on an unbiased test dataset. This demonstrates that our algorithm allows networks trained on biased data to generalize better to unbiased settings, by removing a known bias from the feature representation of the network.

This is a significant step towards trusting that a network definitely isn't basing its decisions on the wrong reasons. With increasing use of neural networks in government, law, and employment to make life-changing decisions, it is of great importance, that undesirable social biases are not encoded in the decision algorithm.

In the context of genetic disorders, datasets are generally small and often all available faces for a syndrome are collected from a single source. This can introduce strong image level biases, e.g. background and lighting conditions, and biases towards the ancestral origin of the population at the image source. We see our algorithm being used to guide the network to ignore these known biases. This is especially useful for our work in Chapter 6, where we create a face embedding that we hope will generalise to new genetic disorders that did not form part of the training data. By removing spurious variations from the face embedding, the remaining variations in the learnt feature representations could be associated more closely to clinically relevant features.

Some spurious variations are easier to remove than others — applying different weights to each spurious variation classifier could account for these differences. A dynamic-weighting scheme, as in (Yin and Liu, 2018), to automatically weight these different tasks at different time-points during training may improve convergence.

6

Interpolated Clinical Face Phenotype Space

Contents

6.1 Datasets 105
6.2 Interpolating Between Faces 108
6.2.1 Sigmoid-Weighted Appearance Blending 109
6.3 Interpolating Between Faces to Improve Classification 112
6.4 Methods for Learning Face Embeddings 113
6.4.1 Siamese Network with Curriculum Learning 113
6.4.2 Interpolating Between Classes with Triplets 114
6.5 Interpolated Face Embeddings 115
6.5.1 Experiments 115
6.5.2 Implementation Details 116
6.5.3 Results and Discussion 117
6.6 Evaluation on DDD Study Dataset 119
6.7 Conclusion 122

This chapter presents a novel image augmentation approach for creating a robust deep face embedding that clusters patients based on shared facial phenotypes. Such an embedding could be used to find groups of patients that share underlying genetic variants. These could then be studied together to characterise the shared genetic disorder. Training such an embedding, however, is a challenging task, as datasets for training only cover a small fraction of all known genetic disorders. If an embedding

is to be useful for clinical use, it would need to generalise beyond syndromes present in the training dataset and encode facial phenotype information more generally.

Ferry et al. (2014) introduced the concept of a "Clinical Face Phenotype Space" by measuring distances in an embedding space that was trained using engineered-feature based representations of images from the LSITW dataset (Section 4.1.1). Patients with shared disorders were shown to lie close to each other in this embedding space. Furthermore, they showed that this embedding was capable of clustering syndromes that were not present in the training data (Ferry et al., 2014).

In Chapter 4, we showed that deep representations outperform engineered-feature based representations for classifying genetic disorders. This motivates us to investigate the use of CNNs to create a deep face embeddings, where we hope to see a similar improvement. Creating a deep genetic disorder embedding, however, presents us with two main challenges. The first, is a lack of training data to learn an accurate syndrome face embedding. And second, the face embedding space between classes is poorly defined, as relationships between syndromes are not captured during training.

Zhang et al. (2017) introduce the idea of "mixing" images from different classes and learning the fractional contribution of each class to the composite. This is achieved by naively blending two images together and creating a label vector that contains a fraction label for each class that is present. If we consider a binary classification task between cats and dogs, cat images may be assigned the one-hot label [1, 0] and dogs the label [0, 1]. A mixture of the two e.g., a cat image with 30% opacity overlayed on a dog image with 70% opacity would be assigned a label of [0.3, 0.7]. This forces the network to learn the inter-class relationships between the two species. The training images, however, do not look realistic when simply overlayed. We mitigate this problem by using landmark guided interpolations, as introduced in Section 3.5. The resulting interpolations look more realistic and contain fewer artifacts in the face (Figure 6.1).

We make use of methods from Chapter 3 to create new realistic-looking faces by interpolating between pairs of faces to augment the training dataset. We name the learned embedding the "*Interpolated* Clinical Face Phenotype Space" (I-CFPS).

Our contributions are two-fold:

Within-class interpolations (WC): We create interpolations between faces that share a class, as a form of image augmentation, to improve deep representation training from small datasets.

Between-class interpolations (BC): We create interpolations between faces of different classes, creating face mixtures of two syndromes. This allows us to explicitly model the space between syndromes. We hypothesise that this may improve the generalisation performance of the deep representation.

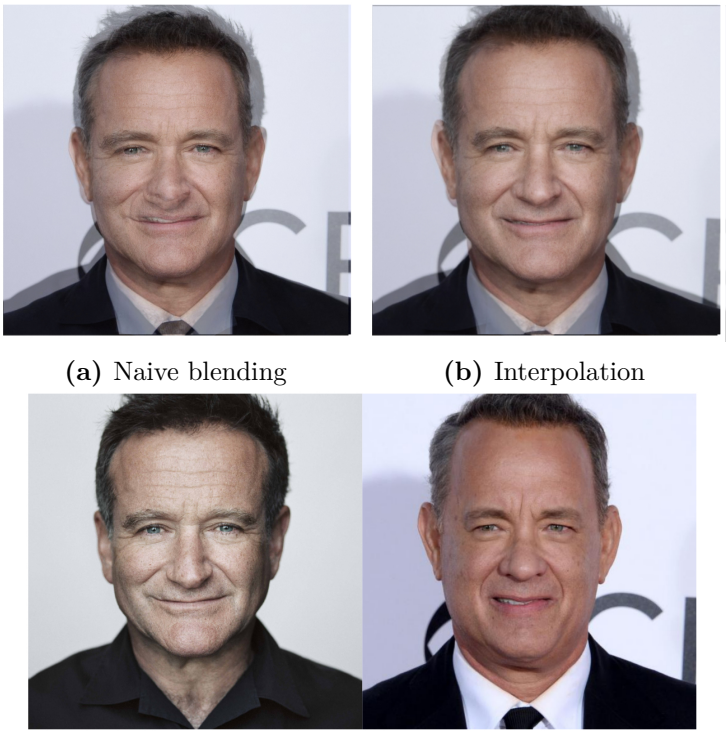
We define three tasks to measure the quality of the learned I-CFPS. K-NN classification accuracies are measured against on three test datasets: 1) test images of known syndromes represented in the training domain, 2) test images of known syndromes from different domains, and 3) test images of unknown syndromes, that were not part of the training domain. Known syndromes refer to class labels seen by the network as part of the training dataset.

Section 6.1 describes the different train and test datasets we used to demonstrate our algorithm. Our method for creating interpolated faces for image augmentation is introduced in Section 6.2. We then investigate its utility as an image augmentation for classification (Section 6.3). Section 6.4 introduces different networks for creating face embeddings, which are demonstrated in Section 6.5. We evaluate our version of the clinical face phenotype space on the DDD dataset in Section 6.6 to measure how well our embeddings generalise to unseen data and conclude in Section 6.7.

6.1 Datasets

Subsets of images from the LSITW, PiPP, and MIR datasets (Section 2.4) were used to train and evaluate the efficacy of using interpolations when training networks in this chapter.

The training subset from LSITW, as described in Section 2.4.1, was used to train the algorithms in this chapter. It will be referred to as LSITW train.



(a) Naive blending **(b)** Interpolation

(c) Input images

Figure 6.1: Differences between (a) Naive blending of two faces, (b) annotation guided interpolation from (c) input images. Annotation guided interpolation yields better quality results. This is a composite image interpolating between two famous celebrities.

The following subsets were used for evaluation:

- **LSITW test** — testing set from LSITW, as described in Section 2.4.1.
- **non-LSITW test** — images from PiPP and MIR of the 8 syndromes that appear in LSITW train.
- **Unseen evaluation** — Images from MIR and PiPP of 38 syndromes that were not part of any training dataset (Table 6.1).

We also made use of the Deciphering Developmental Disorders study (Section 2.4.2) to investigate how well our embeddings generalise to a large clinical dataset of unseen data.

Table 6.1: List of images and syndromes in unseen evaluation subset of MIR and PiPP

Disorder	Images
22q11.2 deletion syndrome	10
Aarskog-Scott syndrome	10
Achondroplasia	10
Acrofacial dysostosis	16
Auriculocondylar syndrome	10
Cardiofaciocutaneous syndrome	33
Cleidocranial dysplasia	12
Coffin-Lowry syndrome	40
Ehlers-Danlos syndrome	26
Goldenhar syndrome	12
Gorlin syndrome	40
Lenz-Majewski hyperostotic dwarfism	18
Mandibulofacial dysostosis	31
Marfan syndrome	50
Melnick-Needles syndrome	12
Monosomy 5p	15
Muenke syndrome	14
Myeloid neoplasms associated with eosinophilia and abnormality of PDGFRA, PDGFRB or FGFR1	10
Noonan syndrome	111
Orofaciodigital syndrome	24
Otopalatodigital syndrome	26
Prader-Willi syndrome	41
Proteus syndrome	15
Rothmund-Thomson syndrome	11
Saethre-Chotzen syndrome	17
Sotos syndrome	52
Stickler syndrome	10
Turner syndrome	17
Waardenburg syndrome	49

6.2 Interpolating Between Faces

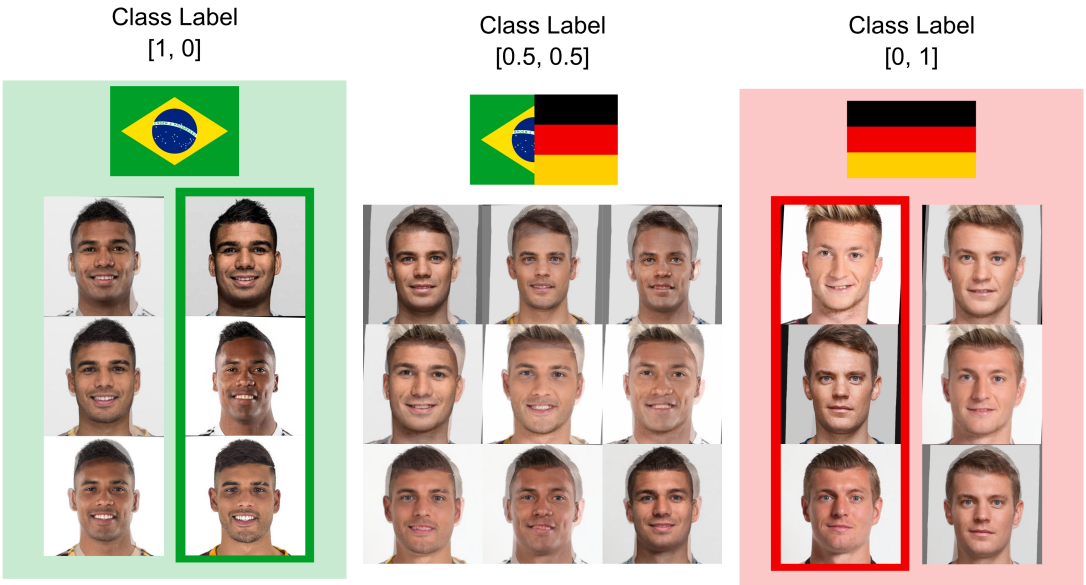
This section introduces the methods for generating interpolations between faces to augment datasets with new, realistic-looking faces.

Interpolations between individual faces are created using the methods described in Section 3.5. Given two images, I_1 and I_2, of labels c_i and c_j the interpolation $I_{1,2}(\gamma)$ is a function of the coefficient of interpolation, γ. This coefficient can range between $0 \leq \gamma \leq 1$ and defines the relative contribution of either image towards the resulting interpolation. When $\gamma = 1$, the resulting interpolated image consists only of I_1, and when $\gamma = 0$, the resulting interpolated image consists only of I_2. The class label of the interpolated image is defined as

$$c_{1,2}(\gamma) = \gamma O(c_1) + (1 - \gamma)O(c_2), \tag{6.1}$$

where $O(c_i)$ is the one-hot representation of class c_i.

Figure 6.2 demonstrates the WC and BC interpolations using several players from German and Brazilian football teams and the resulting ground truth labels.



(a) Brazilian team **(b)** Mixtures **(c)** German team

Figure 6.2: (a) Brazilian football players (left) and WC interpolations (right), (b) BC interpolations, (c) German football players (left) and WC interpolations (right).

6.2.1 Sigmoid-Weighted Appearance Blending

Interpolating between faces can cause unwanted visual artifacts where features do not overlap well. These are especially noticeable outside of the annotated region of the face, e.g. two sets of ears being visible. When creating transition videos, as in Section 3.5, these were not considered problematic. However, these artifacts were found to negatively impact the training of deep networks. We therefore introduce sigmoid-weighted blending, which reduces the appearance of these artifacts in interpolations and, when used to train networks, improves classification performance.

In Section 3.5, we introduced an appearance blending equation, where the same coefficient of interpolation, γ applies to both shape and appearance interpolation (re-stated in Equation 6.2 for convenience). We will refer to this approach as "linear appearance blending" in the remainder of this chapter.

$$A(\gamma) = \gamma A_1 + (1 - \gamma)A_2 \tag{6.2}$$

Sigmoid-weighted interpolation weights appearance and shape differently. The shape equation remains linear, however the appearance coefficient is weighted as follows,

$$A(\gamma) = \sigma(\gamma)A_1 + (1 - \sigma(\gamma))A_2, \tag{6.3}$$

where σ denotes the following sigmoid function:

$$\sigma(\gamma) = \frac{1}{1 + e^{-\epsilon(\gamma - 0.5)}}, \tag{6.4}$$

where ϵ can be used to adjust the shape of the sigmoid function. A value of $\epsilon = 12$ was chosen experimentally and used in our experiments.

Figure 6.3 shows the appearance and shape interpolation coefficients for linear and sigmoid-weighted blending. A more detailed example of both blending modes in practice is shown in Figure 6.4. The differences are subtle and may be more easily observed digitally. At either end of the sigmoid-weighted interpolation ($\gamma < 0.3$ and $\gamma > 0.7$), the changes to the face are analogous to pose and expression augmentations, rather than face mixtures. Towards the middle of the interpolation ($\gamma = 0.5$) both

images contribute equally to the appearance of the interpolation. The artifacts at this point are unavoidable without landmark description for areas outside of the face.

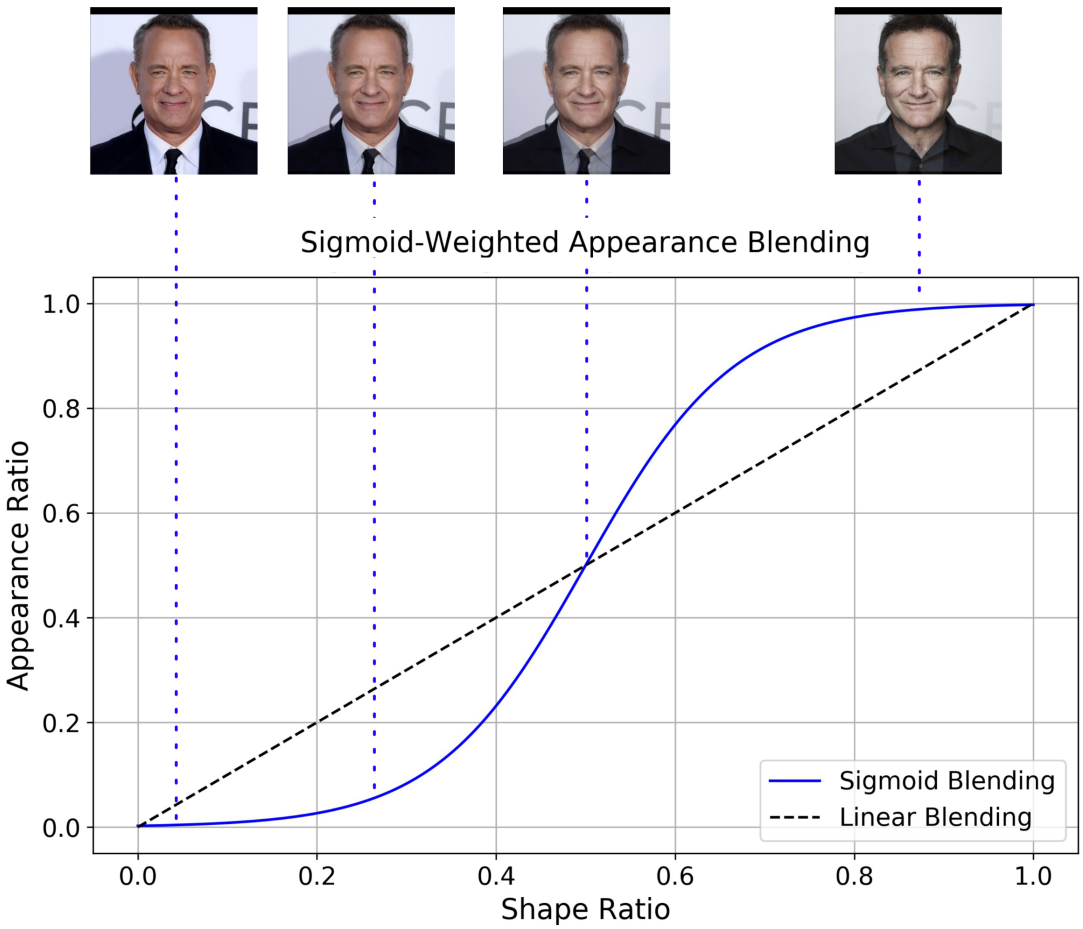
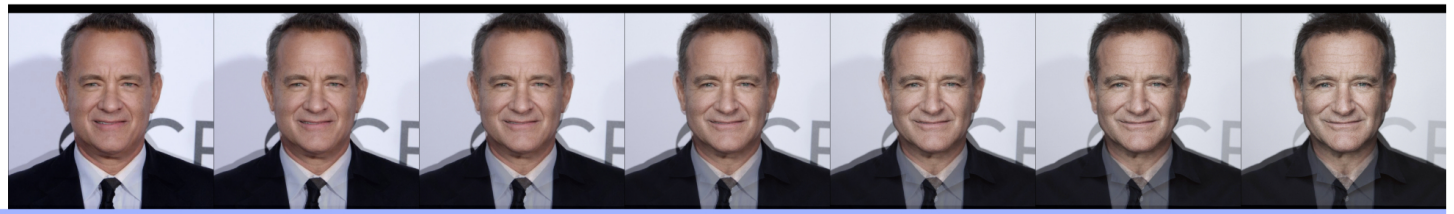
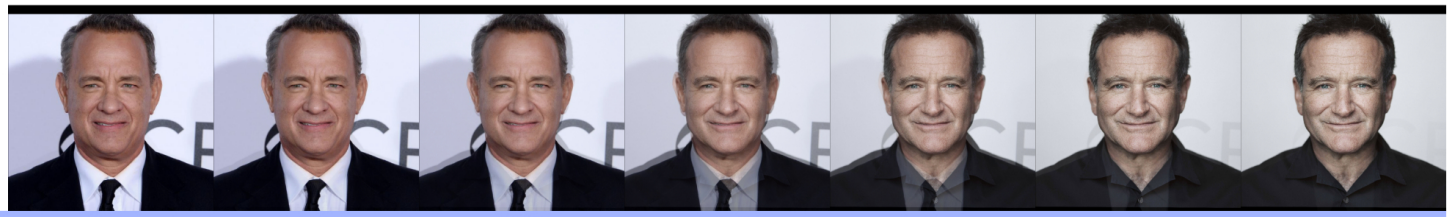


Figure 6.3: The two curves show the shape vs appearance coefficient values for linear (black dotted, γ) and sigmoid-weighted (blue, $\sigma(\gamma)$) blending. The pixel values of one image dominates at either end of the morph ($\gamma < 0.3$ and $\gamma > 0.7$) when using sigmoid-weighted blending. In the middle of the interpolation, ($\gamma = 0.5$) both methods are comparable.



Shape	γ	0.2	0.3	0.4	0.5	0.6	0.7	0.8
Appearance	γ	0.2	0.3	0.4	0.5	0.6	0.7	0.8

(a) Linear appearance blending



Shape	γ	0.2	0.3	0.4	0.5	0.6	0.7	0.8
Appearance	$\sigma(\gamma)$	0.03	0.08	0.23	0.5	0.77	0.92	0.97

(b) Sigmoid-weighted appearance blending

Figure 6.4: Example interpolations using linear and sigmoid-weighted blending. The pixel values of both images are mixed mainly towards the middle of the interpolation with sigmoid-weighted blending.

6.3 Interpolating Between Faces to Improve Classification

We now demonstrate the benefits of WC interpolation in an image classification setting. We used within-class interpolation as a data augmentation method to train networks on the LSITW train data to classify the 8 syndromes.

Pairs of images, $\{I_i, I_j\}$, where $i \neq j$, sharing the same class label were sampled. The interpolation coefficient was sampled at random from a list $\gamma = \{0, 0.1, 0.2, 0.3, ..., 1\}$. Contrary to observations by Zhang et al. (2017), we did not find that restricting $\gamma \in [0.1, 0.4]$ lead to improved classification performance.

We make use of the base architecture of ResNet50 network, pre-trained on VGGFace2 was used (Section 2.4.4) and fine-tune the network on the training dataset of LSITW. The training procedure is detailed in Section 4.2.2. The image augmentations in Section 4.2.4 were applied after interpolating.

Table 6.2 contains classification accuracies on the LSITW test data for networks trained without using interpolation, with linear interpolation, and with sigmoid-weighted interpolation.

Table 6.2: Summary table of average classification accuracies (%) on LSITW test data with different interpolation methods.

	Accuracy (%)
Baseline (No interpolation)	90.6
Linear interpolation	91.6
Sigmoid-weighted interpolation	**92.5**

The baseline network, trained without interpolations, achieves a classification accuracy of 90.6%. This is the same VGG-M network as in Section 4.2.5, that was pre-trained on VGGFace (Section 2.4.3) and fine-tuned to classify the 8 syndromes in the LSITW dataset (Section 2.4.1). Augmenting the training dataset with linear interpolations, the accuracy increases to 92.1% (+1.5%). Sigmoid-weighted interpolation performs best, at 93.0%, a 2.4% improvement upon the baseline network.

This improvement in classification accuracies confirms that interpolating between faces is a valuable image augmentation method for improving deep feature representations. In the remainder of this chapter, we investigate their use in training face embeddings that can generalise to beyond syndromes seen during training.

6.4 Methods for Learning Face Embeddings

This section describes the methods for learning a deep feature representation, which acts as an embedding space in which patients with shared syndromes lie closer to each other than patients with different syndromes. Three approaches were considered: siamese networks with curriculum learning (Section 6.4.1), prototypical networks (Section 2.2.4) and triplet networks (Section 6.4.2).

6.4.1 Siamese Network with Curriculum Learning

We make use of the siamese network approach (Section 2.2.4) to learn embeddings. Curriculum learning (Nagrani et al., 2018) was used to further improve the learned embedding. This method samples progressively harder negative pairs to gradually make the embedding more sensitive to these difficult cases.

When constructing a batch, an equal number of positive and negative pairs were sampled to ensure that the loss function was balanced for both cases. The negative samples were sampled randomly until the loss converged, after which curriculum learning was used to select progressively harder negatives (Nagrani et al., 2018). Positives are sampled randomly throughout the procedure. The distance from an image of class $y = k$ is computed to every negative image of class $y \neq k$ in the training batch and sorted. A difficulty parameter τ, that corresponds to the percentile of the ranked negatives, is used to select an appropriately hard negative: the negative with the smallest distance to the anchor is defined as the hardest negative for which $\tau = 1$ and $\tau = 0$ corresponds to selecting the easiest negative. The value for τ was increased gradually, starting from $\tau = 0.3$ up until $\tau = 0.8$.

6.4.2 Interpolating Between Classes with Triplets

The BC triplet sampling method is based on the triplet network introduced in Section 2.2.4. At each instance, three images are sampled from two classes (c_1, c_2): the anchor image, x, from the class c_1 and two images, x_1, x_2 from class c_2. For each of the second class' images, an interpolation is created between it and the anchor with different interpolation ratios, γ_1, γ_2. The magnitude of the interpolation factors $(\gamma_+ > \gamma_-)$ decides which of the interpolations is chosen as the positive or negative in the triplet (x_+, x_-). The BC triplet sampling approach is shown in Figure 6.5.

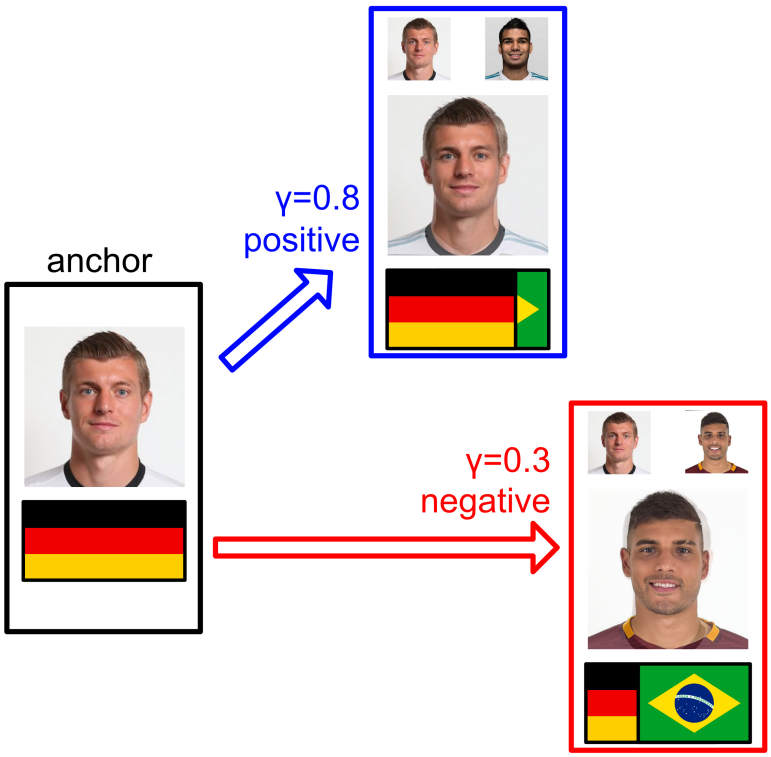


Figure 6.5: BC triplet sampling method: Three images are sampled, an anchor image from class A (German) and two from class B (Brazilian). For each A-B pairing, a random interpolation ratio is sampled and an interpolation is created. The interpolation that is more similar to the anchor is chosen as the positive, and the other image is chosen as the negative in the triplet.

The triplet loss function is shown in Equation 6.5. The aim is to create an embedding space in which x and x_+ are closer to each other than x and x_-

by at least a margin m.

$$L(x, x_+, x_-) = \max(0, m + D(x, x_+) - D(x, x_-)), \tag{6.5}$$

where $D(x_i, x_j)$ denotes the distance between two images x_i, x_j in embedding space.

6.5 Interpolated Face Embeddings

This section presents our experiments and results when using WC and BC interpolations to create deep genetic face embeddings.

6.5.1 Experiments

We trained a ResNet50 network architecture, pre-trained on VGGFace2 (Section 2.4.4) to learn interpolated face embeddings in the following three settings:

- **Class** — Classification network from Section 6.3
- **Siamese** — Siamese network, initialised from Class network
- **Prototypical** — Prototypical network, initialised from Class network

Baseline networks were trained on the LSITW train data without augmenting the training dataset with interpolations. These are compared to networks trained with WC interpolation, networks trained with BC interpolation, and network trained first with WC interpolation, followed by BC interpolation.

We evaluated the clustering performance of the face embeddings in three settings. First the classification accuracies on test data from the same domain were evaluated (LSITW test). Secondly, the generalisation performance to images of seen syndromes in a different domain was measured (non-LSITW test). Third, the generalisation performance to unseen syndromes in a different domain was measured (Unseen evaluation).

Networks trained with WC interpolations will be denoted with the "WC-" prefix, e.g. WC-Class. If the network was fine-tuned further with BC triplet learning (using the loss defined in Equation 6.5), it is denoted with a "+BC" suffix, e.g.

WC-Class+BC denotes the Class network trained with WC interpolations and then fine-tuned with BC triplet learning.

The experiments in the following sections use the sigmoid-weighted interpolation method from Section 6.2.1. The image augmentations in Section 4.2.4 were applied after interpolating.

During testing, an embedding was extracted from a network for each image. The clustering performance of the face embeddings was then measured using K-NN clustering. We used a range of k-values, $k = \{1, 2, 3, 4, 5, .., 10\}$. The k value that resulted in the best classification accuracies was chosen and the accuracies reported. To avoid skewing results in favour of patients with multiple images, each patient could only appear once in any cluster.

6.5.2 Implementation Details

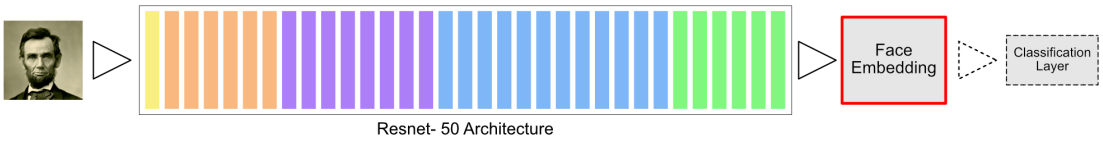


Figure 6.6: Base network architecture. The face embedding layer is a fully connected layer with a 128-dimensional output. To train the classification network, an additional classification layer was appended. This layer was discarded when training the siamese, prototype and triplet embeddings.

The trained classification networks from Section 6.3 were used as the base networks for all following experiments. The classification layer was removed and all losses act on the penultimate layer that we denote as the face embedding layer (Figure 6.6). The learned face embedding is the 128-dimensional output of the fully-connected embedding layer.

The weights in all layers apart from the embedding layer of the network were frozen. Freezing those weights did not adversely affect the performance of the learned embeddings and significantly reduced training times. The loss function is optimised using stochastic gradient descent with a constant learning rate of $L_r = 0.0001$ for the embedding layer.

For each interpolation a random γ was sampled as in Section 6.3.

6.5.3 Results and Discussion

In the following, classification results for the models from Section 6.5 evaluated on LSITW test, non-LSITW test, and Unseen evaluation datasets are presented.

Classification of Seen Syndromes in the Same Domain

This section presents the classification results of our models evaluated on LSITW test. The LSITW test set contains the same 8 syndromes as were seen during training. The patients in the LSITW test set were not, however, seen during training.

WC interpolation improved the baseline network accuracies on LSITW test for every training regime. These are presented in Table 6.3. The best accuracy was achieved with the Class method (92.5%), that used the softmax classification loss for training.

Further fine-tuning with the BC triplet loss reduced accuracies for all categories.

Table 6.3: K-NN classification accuracies (%) of the experiments outlined in Section 6.5 on LSITW test. The accuracy of assigning classes randomly is 12.5%.

LSITW test	Baseline	Baseline+BC	WC	WC+BC
Class	90.6	85.2	**92.5**	88.5
Siamese	87.5	84.0	**89.9**	85.1
Prototypical	87.3	87.1	**92.1**	92.0

Classification of Seen Syndromes in a Different Domain

This section presents classification accuracies of our models evaluated on images in the non-LSITW dataset, which contains images from different datasets of the 8 syndromes seen during training.

When evaluating the trained models on the non-LSITW test dataset, we noticed a significant reduction in classification accuracies. This is most likely due to the domain shift, e.g. images in PiPP are scans from publications. Table 6.4 contains these accuracies across all training regimes.

Again, WC interpolation proves to be a useful augmentation method, improving classification accuracies for all but one method (WC-Siamese+BC). The best result was achieved by the WC-Prototypical model with a classification accuracy of 66.8%, 5.3 times better than assigning a class randomly.

Further fine-tuning with BC triplet learning significantly improved the baseline networks (+3.9 to 5.5%). Further training with BC triplet learning of networks trained with WC interpolations reduced classification accuracies.

Table 6.4: K-NN classification accuracies (%) of the experiments outlined in Section 6.5 on non-LSITW test. The accuracy of assigning classes randomly is 12.5%.

non-LSITW test	Baseline	Baseline+BC	WC	WC+BC
Class	52.1	57.6	**58.2**	58.0
Siamese	56.5	60.4	**60.5**	58.8
Prototypical	56.7	61.4	**66.8**	66.5

Classification of Unseen Syndromes

We finally present classification accuracies of our models evaluated on the Unseen evaluation data that contains syndromes not seen during training. This is to assess the generalisation performance beyond the 8 syndromes in the training dataset.

All models perform significantly better than random chance (3.3%), which suggests that an embedding learned using a small number of syndromes captures clinically relevant features that extend to unseen syndromes.

The embedding learned using the WC-Prototypical+BC regime can classify syndromes over 10 times better (33.5%) than random chance (3.3%). This embedding was also significantly better than the baseline Class regime (+8.9%). This suggests that learning the relationships between classes can improve the generalisation performance of the embedding space to new, unseen syndromes.

Table 6.5: K-NN classification accuracies (%) of the experiments outlined in Section 6.5 on the Unseen evaluation dataset. The accuracy of assigning classes randomly is 3.3%.

Unseen evaluation	Baseline	Baseline+BC	WC	WC+BC
Class	24.6	31.0	27.0	**31.3**
Siamese	31.1	31.1	**32.7**	31.6
Prototypical	31.2	31.7	32.6	**32.9**

The embeddings for images from the Unseen syndromes, computed using the WC-Prototypical+BC networks, are visualised in Figure 6.7. The embeddings were dimensionality-reduced to 50-dimensional vectors using PCA before using t-SNE to map them onto 2 dimensions. Though these groups were not seen during training, images for several of the syndromes form clusters.

6.6 Evaluation on DDD Study Dataset

To assess how our face embeddings would cluster unseen patients, we conducted experiments on the large-scale clinical DDD dataset (Section 2.4.2). We trained a network using the WC-Prototypical+BC method on 16 syndromes from the LSITW, MIR and PiPP datasets. This model performed the best when clustering unseen syndromes, and was thus selected for this experiment. This model will be referred to as the Interpolated Clinical Face Phenotype Space model (I-CFPS). The I-CFPS was used to compute embeddings for the images in the DDD dataset. A single embedding was used for each patient and if multiple images were present for a patient, their embeddings were averaged. We compared the Euclidean distances in embedding space between random pairs of patients and pairs that had genetic variants in common. If patients with shared genetic variants lie closer than random pairs in the I-CFPS embedding does indeed capture clinically relevant phenotype information.

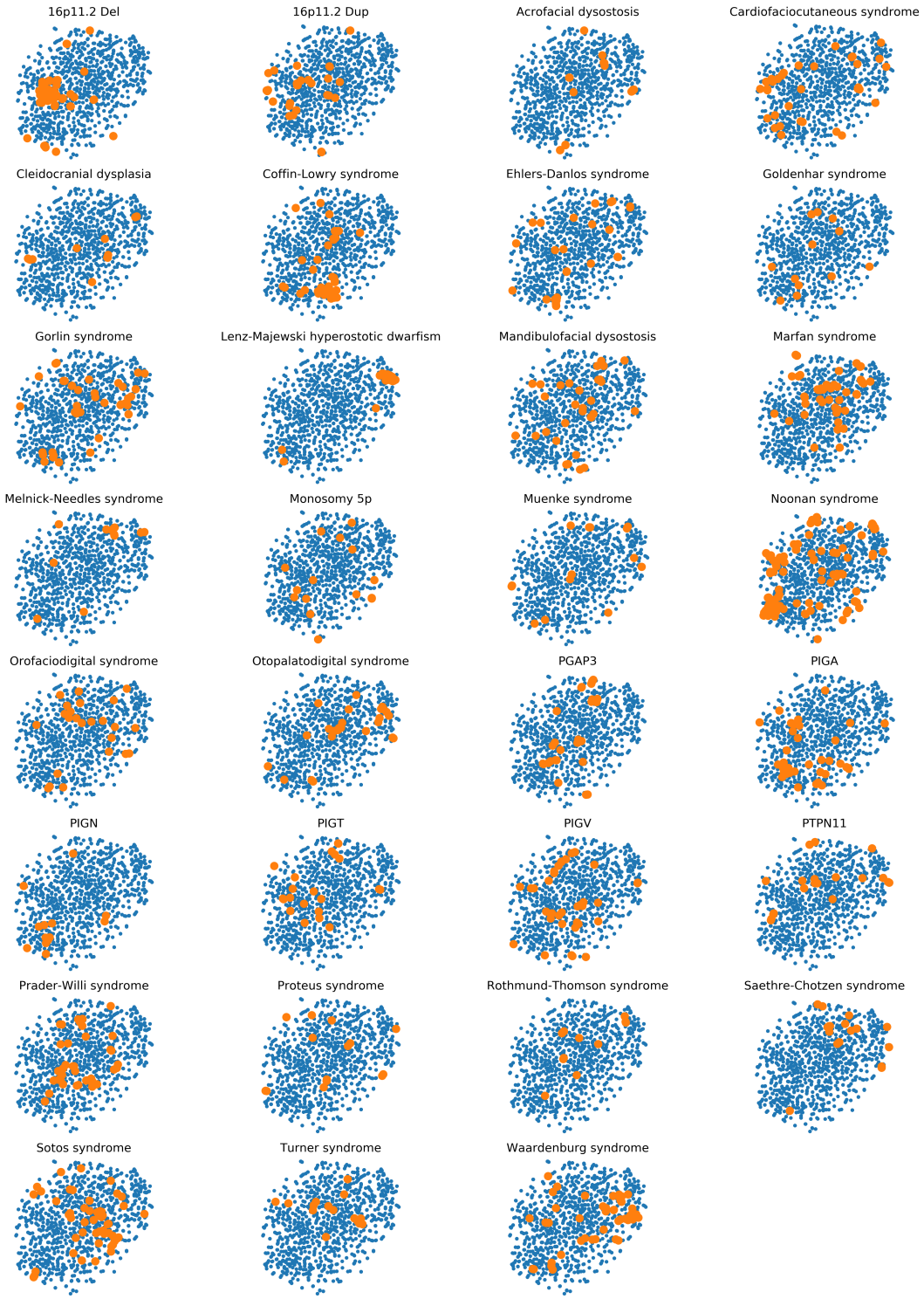


Figure 6.7: t-SNE visualisation of WC-Prototypical+BC embeddings of Unseen evaluation data after PCA to $N = 50$. In each plot, one of the syndrome groups is highlighted (orange) vs the rest of the images in the Unseen evaluation dataset (blue).

Two experiments were conducted to evaluate the clustering performance of the I-CFPS on the DDD study dataset:

1. Distance between patients with the same confirmed diagnoses
2. Distance between patients sharing genetic variants

Distance between patients with the same confirmed diagnoses

As mentioned in Section 6.1, information about a confirmed diagnosis is only available for 135 patients. Of these, there are 8 positive pairs of patients that shared the same diagnosis. The average embedding distances between random pairs of patients was $\mu = 98.6, \sigma^2 = 21.0$. The 8 positive pairs had an average embedding distance of $\mu = 78.7, \sigma^2 = 22.2$. Therefore, it seems that the embedding space captures clinically relevant phenotype information. However, we realise that the sample size is very small to draw significant conclusions.

Distance between patients sharing genetic variants

The embedding distances of pairs of patients selected at random were compared to those patients that shared genetic variants. The distributions are shown Figure 6.8. The value of n denotes the number of patients that share each genetic variant. Most patients sharing genetic variants on average lie closer to each other in embedding space than random pairs.

The Mann-Whitney-U test was conducted to compare each genetic variant distribution to randomly sampled pairs were computed to check for statistical significance. To account for the difference in sample sizes and stochastic variance, this test was repeated 1000 times and the maximum p-value is reported. Eight of the 14 genetic variants had a statistically different ($p < 0.05$) distribution from the random pairs, namely ADNP, DDX3X, DYRK1A, EHMT1, FOXP1, GATAD2B, MED13L, and SYNGAP1. Of these eight variants six have been associated with dismorphic facial features, e.g. DDX3X is associated with the “Abnormal facial shape” HPO term in the Human Phenotype Ontology (Köhler et al., 2019). These

genes were not present in the training dataset, which further confirms that the I-CFPS can generalise to unseen genetic disorders.

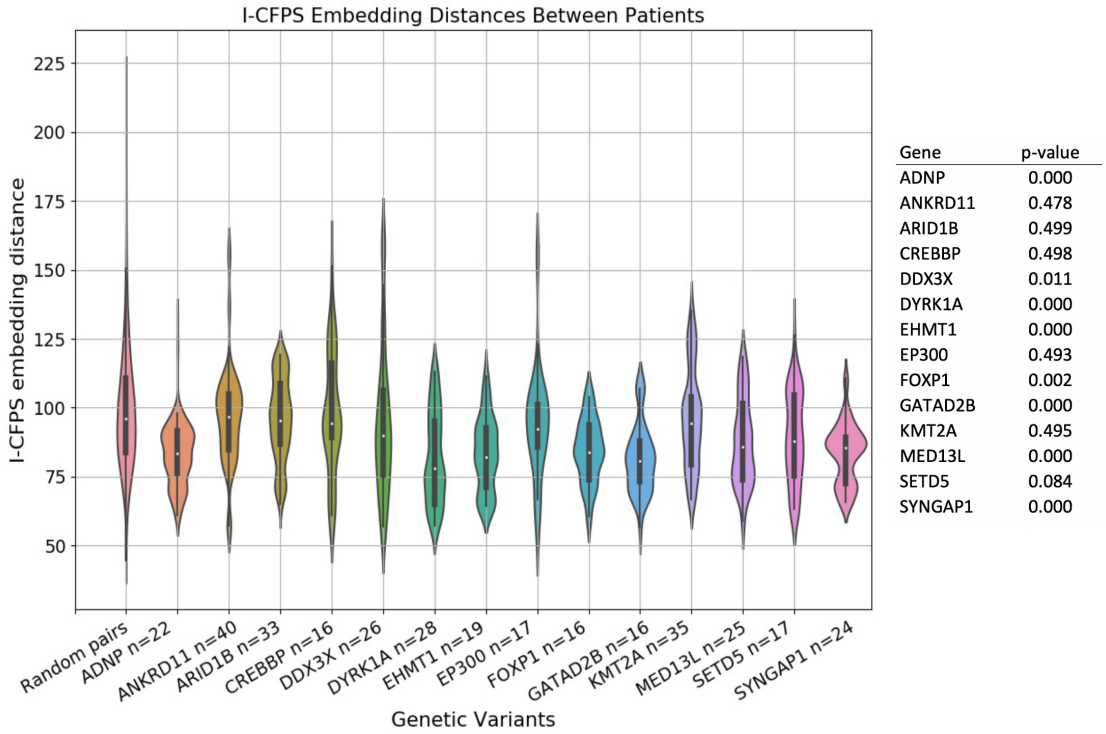


Figure 6.8: Violin plots of distances in I-CFPS embedding space between random pairs of patients and patients sharing genetic variants in DDD dataset. Mann-Whitney-U test p-values between the random pair distribution and genetic variants are shown on the right.

6.7 Conclusion

We investigated the creation of a face embedding, the 'Interpolated Clinical Face Phenotype Space' (I-CFPS), that can cluster patients with shared genetic disorders and generalise to syndromes that were not present in the training dataset. To this end, we introduced an image augmentation method to create new faces by creating interpolations between pairs of faces. The quality of these interpolations was improved by weighting the appearance contribution of either images with a sigmoid function with respect to the shape contribution.

Interpolating between images of the same class (WC interpolation) was shown to improve classification accuracies of a ResNet50 network on the LSITW test

dataset from 90.6% to 92.5%. WC interpolation also improved the generalisation performance of embeddings learned using siamese and prototypical networks. This shows that generating new faces by interpolating between them within a class is a valuable image augmentation. Further training with between-class interpolations with triplet networks was shown to slightly improve the generalisation performance of embeddings in some cases. More research may be necessary on the efficacy of between-class interpolations: understanding what the clinical relationships between pairs of syndromes is could help model these relationships more accurately than assuming these to be linear.

We trained the I-CFPS on images from several clinical datasets of patients with 16 syndromes. These were used to compute embeddings from images in the clinical DDD dataset using our I-CFPS model of disorders that were not seen during training. We showed that patients sharing genetic variants clustered closer than random chance. This confirms that our methods to augment facial datasets with interpolations can help learn face embedding that generalise to unseen syndromes.

7 Conclusion

Contents

7.1 Achievements 124
7.2 Suggestions for Further Work 127

We conclude this thesis by summarising the main findings and implications of our work, as well as suggesting avenues for future research.

7.1 Achievements

Recent developments in deep learning have lead to a plethora of successes in computer vision tasks that had until quite recently been considered too difficult to solve. However, these successes rely to a great extent on large amounts of high quality training data to achieve super-human levels of accuracy on many different classification tasks.

Inspired by these remarkable gains, we explored ways to leverage these deep learning techniques to aid the classification of genetic disorders from facial photographs. This application domain comes with significant challenges that make applying deep learning methods non-trivial due to ethical and legal concerns and strictures. The data available for different genetic disorders is often scarce and

patient privacy is an important concern. Furthermore, there are a large number of genetic disorders, most of which we do not have access to facial images for. There therefore exists a need to create models that generalise beyond syndromes that are present in the training data.

The contributions in this thesis go some way towards leveraging the success of deep learning on generic computer vision tasks to improve the detection of and the research into genetic disorders.

In Chapter 3, we presented methods for the creation of realistic average faces that preserve syndrome phenotypes, whilst de-identifying constituent individuals. We showed that facial asymmetry could be preserved using a combination of shape measures. Exaggerated averages and transitions between healthy controls and syndromes averages were used to highlight clinically relevant phenotypes. These averages were used in several publications (Deciphering Developmental Disorders Study et al., 2017; Bengani et al., 2017; Reijnders et al., 2018a,b), where visualising the commonalities between patients was a valuable output. Due to patient consent restrictions, images of individual patients would not have been permitted to be shared.

In Chapter 4, we contrasted several engineered-feature based and deep representations for the classification of genetic disorders. First, we investigated improvements upon engineered-feature based descriptors from Ferry et al. (2014). The shape descriptor was improved by improving the accuracy of automatic landmark placement on the face using SDM. The appearance descriptor was improved upon using Fisher vectors to densely describe the face. While both of our improvements outperform previous work, the joint use of shape and appearance descriptors with late data fusion was shown to perform better than individual descriptors. Lastly, we adapted deep neural networks to our classification task using transfer learning. The VGG-M, ResNet50, and SENet50 networks were compared. SENet50, pre-trained on VGGFace2, and fine-tuned on the training dataset of LSITW outperformed

all engineered-feature based methods significantly. This shows that even smaller datasets can leverage the power of deep learning.

In Chapter 5, presented a novel approach for removing multiple known dataset biases or spurious variations from a deep neural network. We demonstrated our algorithm's efficacy on face classification tasks of age, ancestral origin, pose and gender. The resulting feature representations remained informative for one task, whilst simultaneously being uninformative for the others. We demonstrated that our algorithm allows networks trained on biased data to generalize better to unbiased settings, by removing known biases from the feature representation of the network. Our algorithm can make use of multiple datasets, each describing a source of variation (e.g. age, gender, ancestral origin) that needs to be removed from a feature representation, trained for a primary task (e.g. detecting genetic disorders). This work can be used to train CNNs on small datasets or those that contain biases, and control which factors should not influence their feature representation.

In Chapter 6, we introduced an image augmentation method with two different types of interpolation between faces. First, we interpolated between images within a class. This was shown to improve classification accuracies of a ResNet50 network on the LSITW test dataset from 90.6% to 92.5%. Training with sigmoid-weighting of the appearance contribution of either image with respect to the shape contribution resulted in better classifications. We then conducted further fine-tuning with between-class interpolations with triplet networks. These were shown to improve the generalisation performance of embeddings learned using classification or prototypical networks. We developed a novel face embedding for encoding phenotype information, the Interpolated Clinical Face Phenotype Space (I-CFPS), which was trained on patients images for 16 syndromes from several clinical datasets. The I-CFPS was used to compute embeddings for images in the clinical DDD dataset of disorders that were not seen during training. We showed that patients sharing genetic variants clustered closer than random chance. This confirmed our hypothesis, that face

interpolations within and between classes can be used to improve generalisation beyond seen syndromes.

7.2 Suggestions for Further Work

There are a number of exciting avenues of research that could constitute future research directions.

Profile faces and including the cranium. The research in this thesis was focused on the use of frontal faces. Profile faces, however, also contain significant information about genetic disorders that cannot be captured using 2-dimensional frontal images. An extension of our work would be to procure a dataset that includes profile faces of individuals with genetic disorders. This could then be used to create profile face averages with a profile face landmark annotation model. Currently, the cranium cannot be accurately averaged with our averaging algorithm, as we did not have an annotation algorithm with cranial landmarks. As changes in cranium shape and size can be clinically relevant, training a landmark annotator that includes landmarks on the cranium could result in more informative average faces. Profile images could also be added as an additional set of face parts to the part-based classification approach in Chapter 4.

3-dimensional images. The cost of 3-dimensional photography is constantly reducing, with many consumer smart-phones now containing depth sensors and multiple cameras. This adds a wealth of data that can be used to better describe the face, which, is not possible from frontal 2-dimensional imagery. This makes the case for developing algorithms for "ordinary" 3-dimensional imagery, that are not captured under controlled settings. The algorithms in this thesis can be extended to use 3-dimensional imagery, e.g. to create 3-dimensional average faces using x,y,z coordinates or 3-dimensional CNNs for deep feature extraction.

Generating deep averages. Our average faces algorithm (Chapter 3) currently relies on geometric transformation in the image space. There have been significant advancements in the field of generative adversarial networks (GAN) for the purposes of generating high-quality faces (Karras et al., 2018). Latent space perturbations in the CNN could be used to create faces that retain syndrome characteristics, but new, fake identities. Applying attributes by varying latent space variables, as in Choi et al. (2017); Gou et al. (2019) could also be used to create transitions between disorders, or to apply the facial characteristics of certain disorders on control faces to visualise their effects. However, existing work relies on large datasets and the lack of training data in the domain of genetic disorders will prove to be a significant challenge.

Ensemble models. The outputs of each part-based model from Chapter 4 are currently weighted equally. This could be extended by training ensemble models (Sharkey, 1999), where all face-part networks can be learned together, to take into account the different strengths and weaknesses for each face part in classifying certain syndromes. This even allows fo the use of multiple architectures in an ensemble (Minetto et al., 2018). The I-CFPS embedding training, introduced in Chapter 6, could then be built by combining the deep face embeddings of each face part. A fully connected layer could be used to learn the optimal fusion between each face part embedding to create a single face embedding.

Removing biases for genetic disorder classification. We envisage our bias removal algorithm (Chapter 5) being used to mitigate the effects of dataset bias in clinical face datasets. A number of spurious variations can affect the feature representations learned by a CNN. For example, a group of patients photographed at a hospital in northern Europe may all share a single ancestral origin. The removal of this signal using our LAOFIW dataset (Section 5.2.1) could improve the generalisation performance of genetic face classifiers to other ancestral origins. Similarly, if multiple related individuals are present in a small dataset for a disorder,

the kinship signal may incorrectly be interpreted to be useful to characterise the phenotype of the disorder. The use of our bias removal algorithm with a kinship database, e.g. Dawson et al. (2018), constitutes an interesting approach to mitigate this problem. We would also like to investigate the effects of removing age and gender signals from classifiers using the cleaned IMDB dataset (Section 5.2.2). Additionally, the use of domain confusion (Tzeng et al., 2015) to make networks agnostic to which dataset an image was drawn from may mitigate the effects of environment factors, e.g. background or lighting conditions, to bias the feature representation of a CNN. Visualising which part of the image was used for classification, as in Burns et al. (2018), could be used to assess whether the bias removal was successful.

The effect of unlearning these spurious variations to improve the I-CFPS could also be investigated. The results could be compared to that of the networks that were trained without explicit bias removal (Section 2.4.2), to investigate whether generalisation to unseen syndromes would improve.

Using interpolations to create robust face templates. Combining feature representations of multiple images of individuals has been used to reduce the effects of inter-image variability (Klare et al., 2015). As an extension, our interpolation-based image augmentation method in Section 6 could also be used at test time to create average feature representations. Instead of averaging the feature representations of several disjointed images, the space between them could be interpolated. The resulting average face representation may be more robust to noise.

Clinical translation. The work in this thesis shows that there is significant potential to make use of computer vision algorithms to aid clinicians in the facial detection of genetic disorders. When clinicians narrow down diagnosis to several known syndromes, our classifiers can be used to identify which one is most likely present. However, more work needs to be done to curate more comprehensive datasets. These would allow for more classification models that can detect more syndromes, and a more generalised I-CFPS, to help find patients that share clinically

relevant phenotypes. The predictions from our models can not yet be used as clear diagnoses, but may serve as guidelines, to support clinicians.

Bibliography

Timo Ahonen, Abdenour Hadid, and Matti Pietikäinen. Face Recognition with Local Binary Patterns. In Tomás Pajdla and Jiŗí Matas, editors, *Computer Vision - ECCV 2004*, Lecture Notes in Computer Science, pages 469–481. Springer Berlin Heidelberg, 2004. ISBN 978-3-540-24670-1.

Joanna S. Amberger, Carol A. Bocchini, François Schiettecatte, Alan F. Scott, and Ada Hamosh. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Research*, 43(D1): D789–D798, January 2015. ISSN 0305-1048.

P a Baird, T W Anderson, H B Newcombe, and R B Lowry. Genetic disorders in children and young adults: a population study. *American journal of human genetics*, 42(5):677–93, may 1988. ISSN 0002-9297.

Brunilda Balliu, Rolf P. Würtz, Bernhard Horsthemke, Dagmar Wieczorek, and Stefan Böhringer. Classification and Visualization Based on Derived Image Features: Application to Genetic Syndromes. *PLOS ONE*, 9(11):e109033, November 2014. ISSN 1932-6203.

L. Basel-Vanagaite, L. Wolf, M. Orin, L. Larizza, C. Gervasini, I. D. Krantz, and M. A. Deardoff. Recognition of the Cornelia de Lange syndrome phenotype with facial dysmorphology novel analysis. *Clinical Genetics*, 89(5):557–563, 2016. ISSN 1399-0004.

Peter Belhumeur, David W. Jacobs, David Kriegman, and Neeraj Kumar. Localizing parts of faces using a consensus of exemplars. In *The 24th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 545–552, June 2011.

Hemant Bengani, Mark Handley, Mohsan Alvi, Rita Ibitoye, Melissa Lees, Sally Ann Lynch, Wayne Lam, Madeleine Fannemel, Ann Nordgren, H. Malmgren, M. Kvarnung, Sarju Mehta, Shane McKee, Margo Whiteford, Fiona Stewart, Fiona Connell, Jill Clayton-Smith, Sahar Mansour, Shehla Mohammed, Alan Fryer, Jenny Morton, Detelina Grozeva, Tara Asam, David Moore, Alejandro Sifrim, Jeremy McRae, Matthew E. Hurles, Helen V. Firth, F. Lucy Raymond, Usha Kini, Christoffer Nellåker, D. D. D. Study, and David R. FitzPatrick. Clinical and molecular consequences of disease-associated de novo mutations in SATB2. *Genetics in Medicine*, 19(8):900–908, August 2017. ISSN 1530-0366.

Kym M. Boycott, Ana Rath, Jessica X. Chong, Taila Hartley, Fowzan S. Alkuraya, Gareth Baynam, Anthony J. Brookes, Michael Brudno, Angel Carracedo, Johan T. den Dunnen, Stephanie O. M. Dyke, Xavier Estivill, Jack Goldblatt, Catherine Gonthier, Stephen C. Groft, Ivo Gut, Ada Hamosh, Philip Hieter, Sophie Höhn, Matthew E. Hurles, Petra Kaufmann, Bartha M. Knoppers, Jeffrey P. Krischer, Milan Macek, Gert

Matthijs, Annie Olry, Samantha Parker, Justin Paschall, Anthony A. Philippakis, Heidi L. Rehm, Peter N. Robinson, Pak-Chung Sham, Rumen Stefanov, Domenica Taruscio, Divya Unni, Megan R. Vanstone, Feng Zhang, Han Brunner, Michael J. Bamshad, and Hanns Lochmüller. International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases. *The American Journal of Human Genetics*, 100(5): 695–705, May 2017. ISSN 0002-9297.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, New York, NY, USA, 23–24 Feb 2018. PMLR.

Kaylee Burns, Lisa Anne Hendricks, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also Snowboard: Overcoming Bias in Captioning Models. *arXiv:1803.09797 [cs]*, March 2018. arXiv: 1803.09797.

Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. VGGFace2: A Dataset for Recognising Faces across Pose and Age. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 67–74, May 2018.

Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the Devil in the Details: Delving Deep into Convolutional Nets. pages 6.1–6.12. British Machine Vision Association, 2014. ISBN 978-1-901725-52-0.

Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. *arXiv:1711.09020 [cs]*, November 2017. arXiv: 1711.09020.

S. Chopra, R. Hadsell, and Y. LeCun. Learning a Similarity Metric Discriminatively, with Application to Face Verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546, San Diego, CA, USA, 2005. IEEE.

Peter Claes, Jasmien Roosenboom, Julie D. White, Tomek Swigut, Dzemila Sero, Jiarui Li, Myoung Keun Lee, Arslan Zaidi, Brooke C. Mattern, Corey Liebowitz, Laurel Pearson, Tomás González, Elizabeth J. Leslie, Jenna C. Carlson, Ekaterina Orlova, Paul Suetens, Dirk Vandermeulen, Eleanor Feingold, Mary L. Marazita, John R. Shaffer, Joanna Wysocka, Mark D. Shriver, and Seth M. Weinberg. Genome-wide mapping of global-to-local genetic effects on human facial shape. *Nature Genetics*, 50 (3):414–423, March 2018. ISSN 1546-1718.

T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In Hans Burkhardt and Bernd Neumann, editors, *Computer Vision — ECCV'98*, Lecture Notes in Computer Science, pages 484–498. Springer Berlin Heidelberg, 1998. ISBN 978-3-540-69235-5.

David Cristinacce and Tim Cootes. Automatic Feature Localisation with Constrained Local Models. *Pattern Recogn.*, 41(10):3054–3067, October 2008. ISSN 0031-3203.

N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1, June 2005.

Mitchell Dawson, Andrew Zisserman, and Christoffer Nell\a aker. Mining Faces from Biomedical Literature Using Deep Learning. In *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics*, ACM-BCB '17, pages 562–567, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4722-8. event-place: Boston, Massachusetts, USA.

Mitchell Dawson, Andrew Zisserman, and Christoffer Nellåker. From Same Photo: Cheating on Visual Kinship Challenges. *arXiv:1809.06200 [cs]*, September 2018. arXiv: 1809.06200.

Deciphering Developmental Disorders Study, Jeremy F. McRae, Stephen Clayton, Tomas W. Fitzgerald, Joanna Kaplanis, Elena Prigmore, Diana Rajan, Alejandro Sifrim, Stuart Aitken, Nadia Akawi, Mohsan Alvi, Kirsty Ambridge, Daniel M. Barrett, Tanya Bayzetinova, Philip Jones, Wendy D. Jones, Daniel King, Netravathi Krishnappa, Laura E. Mason, Tarjinder Singh, Adrian R. Tivey, Munaza Ahmed, Uruj Anjum, Hayley Archer, Ruth Armstrong, Jana Awada, Meena Balasubramanian, Siddharth Banka, Diana Baralle, Angela Barnicoat, Paul Batstone, David Baty, Chris Bennett, Jonathan Berg, Birgitta Bernhard, A. Paul Bevan, Maria Bitner-Glindzicz, Edward Blair, Moira Blyth, David Bohanna, Louise Bourdon, David Bourn, Lisa Bradley, Angela Brady, Simon Brent, Carole Brewer, Kate Brunstrom, David J. Bunyan, John Burn, Natalie Canham, Bruce Castle, Kate Chandler, Elena Chatzimichali, Deirdre Cilliers, Angus Clarke, Susan Clasper, Jill Clayton-Smith, Virginia Clowes, Andrea Coates, Trevor Cole, Irina Colgiu, Amanda Collins, Morag N. Collinson, Fiona Connell, Nicola Cooper, Helen Cox, Lara Cresswell, Gareth Cross, Yanick Crow, Mariella D'Alessandro, Tabib Dabir, Rosemarie Davidson, Sally Davies, Dylan de Vries, John Dean, Charu Deshpande, Gemma Devlin, Abhijit Dixit, Angus Dobbie, Alan Donaldson, Dian Donnai, Deirdre Donnelly, Carina Donnelly, Angela Douglas, Sofia Douzgou, Alexis Duncan, Jacqueline Eason, Sian Ellard, Ian Ellis, Frances Elmslie, Karenza Evans, Sarah Everest, Tina Fendick, Richard Fisher, Frances Flinter, Nicola Foulds, Andrew Fry, Alan Fryer, Carol Gardiner, Lorraine Gaunt, Neeti Ghali, Richard Gibbons, Harinder Gill, Judith Goodship, David Goudie, Emma Gray, Andrew Green, Philip Greene, Lynn Greenhalgh, Susan Gribble, Rachel Harrison, Lucy Harrison, Victoria Harrison, Rose Hawkins, Liu He, Stephen Hellens, Alex Henderson, Sarah Hewitt, Lucy Hildyard, Emma Hobson, Simon Holden, Muriel Holder, Susan Holder, Georgina Hollingsworth, Tessa Homfray, Mervyn Humphreys, Jane Hurst, Ben Hutton, Stuart Ingram, Melita Irving, Lily Islam, Andrew Jackson, Joanna Jarvis, Lucy Jenkins, Diana Johnson, Elizabeth Jones, Dragana Josifova, Shelagh Joss, Beckie Kaemba, Sandra Kazembe, Rosemary Kelsell, Bronwyn Kerr, Helen Kingston, Usha Kini, Esther Kinning, Gail Kirby, Claire Kirk, Emma Kivuva, Alison Kraus, Dhavendra Kumar, V. K. Ajith Kumar, Katherine Lachlan, Wayne Lam, Anne Lampe, Caroline Langman, Melissa Lees, Derek Lim, Cheryl Longman, Gordon Lowther, Sally A. Lynch, Alex Magee, Eddy Maher, Alison Male, Sahar Mansour, Karen Marks, Katherine Martin, Una Maye, Emma McCann, Vivienne McConnell, Meriel McEntagart, Ruth McGowan, Kirsten McKay, Shane McKee, Dominic J. McMullan, Susan McNerlan, Catherine McWilliam, Sarju Mehta, Kay Metcalfe, Anna Middleton, Zosia

Miedzybrodzka, Emma Miles, Shehla Mohammed, Tara Montgomery, David Moore, Sian Morgan, Jenny Morton, Hood Mugalaasi, Victoria Murday, Helen Murphy, Swati Naik, Andrea Nemeth, Louise Nevitt, Ruth Newbury-Ecob, Andrew Norman, Rosie O'Shea, Caroline Ogilvie, Kai-Ren Ong, Soo-Mi Park, Michael J. Parker, Chirag Patel, Joan Paterson, Stewart Payne, Daniel Perrett, Julie Phipps, Daniela T. Pilz, Martin Pollard, Caroline Pottinger, Joanna Poulton, Norman Pratt, Katrina Prescott, Sue Price, Abigail Pridham, Annie Procter, Hellen Purnell, Oliver Quarrell, Nicola Ragge, Raheleh Rahbari, Josh Randall, Julia Rankin, Lucy Raymond, Debbie Rice, Leema Robert, Eileen Roberts, Jonathan Roberts, Paul Roberts, Gillian Roberts, Alison Ross, Elisabeth Rosser, Anand Saggar, Shalaka Samant, Julian Sampson, Richard Sandford, Ajoy Sarkar, Susann Schweiger, Richard Scott, Ingrid Scurr, Ann Selby, Anneke Seller, Cheryl Sequeira, Nora Shannon, Saba Sharif, Charles Shaw-Smith, Emma Shearing, Debbie Shears, Eamonn Sheridan, Ingrid Simonic, Roldan Singzon, Zara Skitt, Audrey Smith, Kath Smith, Sarah Smithson, Linda Sneddon, Miranda Splitt, Miranda Squires, Fiona Stewart, Helen Stewart, Volker Straub, Mohnish Suri, Vivienne Sutton, Ganesh Jawahar Swaminathan, Elizabeth Sweeney, Kate Tatton-Brown, Cat Taylor, Rohan Taylor, Mark Tein, I. Karen Temple, Jenny Thomson, Marc Tischkowitz, Susan Tomkins, Audrey Torokwa, Becky Treacy, Claire Turner, Peter Turnpenny, Carolyn Tysoe, Anthony Vandersteen, Vinod Varghese, Pradeep Vasudevan, Parthiban Vijayarangakannan, Julie Vogt, Emma Wakeling, Sarah Wallwark, Jonathon Waters, Astrid Weber, Diana Wellesley, Margo Whiteford, Sara Widaa, Sarah Wilcox, Emily Wilkinson, Denise Williams, Nicola Williams, Louise Wilson, Geoff Woods, Christopher Wragg, Michael Wright, Laura Yates, Michael Yau, Chris Nellåker, Michael Parker, Helen V. Firth, Caroline F. Wright, David R. FitzPatrick, Jeffrey C. Barrett, and Matthew E. Hurles. Prevalence and architecture of *de novo* mutations in developmental disorders. *Nature*, 542(7642):433–438, February 2017. ISSN 1476-4687.

Jun Deng, Wei Dong, Richard Socher, Li-Jia Li, Kuntai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

Jelena Djordjevic, Alexei I. Zhurov, and Stephen Richmond. Genetic and Environmental Contributions to Facial Morphological Variation: A 3d Population-Based Twin Study. *PLoS ONE*, 11(9), September 2016. ISSN 1932-6203.

G. J. Edwards, C. J. Taylor, and T. F. Cootes. Interpreting face images using active appearance models. In *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 300–305, April 1998.

H Erhan and J Belotserkovsky. B-23neuropsychological impact of turner syndrome: Importance of parent education, early detection, and intervention: A case study. *Archives of clinical neuropsychology: the official journal of the National Academy of Neuropsychologists*, 29(6):544–545, 2014.

Mark Everingham, Josef Sivic, and Andrew Zisserman. Taking the bite out of automated naming of characters in TV video. *Image and Vision Computing*, 27(5):545–559, apr 2009. ISSN 02628856.

Quentin Ferry, Julia Steinberg, Caleb Webber, David R FitzPatrick, Chris P Ponting, Andrew Zisserman, and Christoffer Nellåker. Diagnostically relevant facial gestalt information from ordinary photos. *eLife*, 3, 2014. ISSN 2050-084X.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-Adversarial Training of Neural Networks. In Gabriela Csurka, editor, *Domain Adaptation in Computer Vision Applications*, pages 189–209. Springer International Publishing, Cham, 2017. ISBN 978-3-319-58346-4 978-3-319-58347-1.

R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *ArXiv e-prints*, November 2013.

Duoduo Gou, Shaolin Zhang, Xin Ning, and Haiying Wang. A Face Aging Network Based on Conditional Cycle Loss and The Principle of Homology Continuity. *2019 International Conference on High Performance Big Data and Intelligent Systems (HPBD&IS)*, pages 264–268, 2019.

A. Gretton, AJ. Smola, J. Huang, M. Schmittfull, KM. Borgwardt, and B. Schölkopf. Covariate shift and local learning by distribution matching. In *Dataset Shift in Machine Learning*, pages 131–160. Biologische Kybernetik, Cambridge, MA, USA, 2009.

Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: Challenge of recognizing one million celebrities in the real world. 02 2016.

Yaron Gurovich, Yair Hanani, Omri Bar, Nicole Fleischer, Dekel Gelbman, Lina Basel-Salmon, Peter Krawitz, Susanne B. Kamphausen, Martin Zenker, Lynne M. Bird, and Karen W. Gripp. DeepGestalt - Identifying Rare Genetic Syndromes Using Deep Learning. *arXiv:1801.07637 [cs]*, January 2018. arXiv: 1801.07637.

Yaron Gurovich, Yair Hanani, Omri Bar, Guy Nadav, Nicole Fleischer, Dekel Gelbman, Lina Basel-Salmon, Peter M. Krawitz, Susanne B. Kamphausen, Martin Zenker, Lynne M. Bird, and Karen W. Gripp. Identifying facial phenotypes of genetic disorders using deep learning. *Nature Medicine*, 25(1):60–64, January 2019. ISSN 1078-8956, 1546-170X.

Peter Hammond, Tim J. Hutton, Judith E. Allanson, Bernard Buxton, Linda E. Campbell, Jill Clayton-Smith, Dian Donnai, Annette Karmiloff-Smith, Kay Metcalfe, Kieran C. Murphy, Michael Patton, Barbara Pober, Katrina Prescott, Pete Scambler, Adam Shaw, Ann C. M. Smith, Angela F. Stevens, I. Karen Temple, Raoul Hennekam, and May Tassabehji. Discriminating Power of Localized Three-Dimensional Facial Morphology. *American Journal of Human Genetics*, 77(6):999–1010, December 2005. ISSN 0002-9297.

Peter Hammond, Michael Suttie, Raoul C. Hennekam, Judith Allanson, Eileen M. Shore, and Frederick S. Kaplan. The Face Signature of Fibrodysplasia Ossificans Progressiva. *American Journal of Medical Genetics. Part a*, 158A(6):1368–1380, June 2012. ISSN 1552-4825.

T C Hart and P S Hart. Genetic studies of craniofacial anomalies: clinical implications and applications. *Orthodontics & craniofacial research*, 12(3):212–20, aug 2009. ISSN 1601-6343.

Kaiming He and Jian Sun. Convolutional Neural Networks at Constrained Time Cost. *arXiv:1412.1710 [cs]*, December 2014. arXiv: 1412.1710.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *arXiv:1512.03385 [cs]*, December 2015. arXiv: 1512.03385.

R Herpers, H Rodax, and G Sommer. A Neural Network Identifies Face with Morphological Syndromes. In *Volume 10: Artificial Intelligence in Medicine*, volume 10 of *Studies in Health Technology and Informatics*, pages 481 – 485. IOS Press, 1993.

Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-Excitation Networks. *arXiv:1709.01507 [cs]*, September 2017. arXiv: 1709.01507.

Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

Aishwariya Iyengar, Rutvij Kulkarni, and T. N. C. Vidya. Koinophilia and human facial attractiveness. *Resonance*, 20(4):311–319, April 2015. ISSN 0973-712X.

L. Karlinsky, M. Dinerstein, D. Harari, and S. Ullman. The chains model for detecting parts by their context. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 25–32, June 2010.

Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. *arXiv:1812.04948 [cs, stat]*, December 2018. arXiv: 1812.04948.

Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, 2014. ISSN 10636919.

Ira Kemelmacher, Steve Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. 12 2015.

Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

Brendan F. Klare, Ben Klein, Emma Taborsky, Austin Blanton, Jordan Cheney, Kristen Allen, Patrick Grother, Alan Mah, Mark Burge, and Anil K. Jain. Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1931–1939, Boston, MA, USA, June 2015. IEEE. ISBN 978-1-4673-6964-0.

Adam Kortylewski, Bernhard Egger, Andreas Schneider, Thomas Gerig, Andreas Morel-Forster, and Thomas Vetter. Analyzing and Reducing the Damage of Dataset Bias to Face Recognition With Synthetic Data. In *CVPR 2019*, 2019.

Martin Kostinger, Paul Wohlhart, Peter M. Roth, and Horst Bischof. Annotated Facial Landmarks in the Wild: A large-scale, real-world database for facial landmark localization. pages 2144–2151. IEEE, November 2011. ISBN 978-1-4673-0063-6 978-1-4673-0062-9 978-1-4673-0061-2.

Robin S. S. Kramer, Rob Jenkins, and A. Mike Burton. InterFace: A software package for face image warping, averaging, and principal components analysis. *Behavior Research Methods*, 49(6):2002–2011, December 2017. ISSN 1554-3528.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

Paul Kruszka, Yonit A. Addissie, Daniel E. McGinn, Antonio R. Porras, Elijah Biggs, Matthew Share, T. Blaine Crowley, Brian H. Y. Chung, Gary T. K. Mok, Christopher C. Y. Mak, Premala Muthukumarasamy, Meow-Keong Thong, Nirmala D. Sirisena, Vajira H. W. Dissanayake, C. Sampath Paththinige, L. B. Lahiru Prabodha, Rupesh Mishra, Vorasuk Shotelersuk, Ekanem Nsikak Ekure, Ogochukwu Jidechukwu Sokunbi, Nnenna Kalu, Carlos R. Ferreira, Jordann-Mishael Duncan, Siddaramappa Jagdish Patil, Kelly L. Jones, Julie D. Kaplan, Omar A. Abdul-Rahman, Annette Uwineza, Leon Mutesa, Angélica Moresco, María Gabriela Obregon, Antonio Richieri-Costa, Vera L. Gil-da-Silva-Lopes, Adebowale A. Adeyemo, Marshall Summar, Elaine H. Zackai, Donna M. McDonald-McGinn, Marius George Linguraru, and Maximilian Muenke. 22q11.2 deletion syndrome in diverse populations. *American Journal of Medical Genetics Part A*, 173(4):879–888, 2017a. ISSN 1552-4833.

Paul Kruszka, Antonio R. Porras, Yonit A. Addissie, Angélica Moresco, Sofia Medrano, Gary T. K. Mok, Gordon K. C. Leung, Cedrik Tekendo-Ngongang, Annette Uwineza, Meow-Keong Thong, Premala Muthukumarasamy, Engela Honey, Ekanem N. Ekure, Ogochukwu J. Sokunbi, Nnenna Kalu, Kelly L. Jones, Julie D. Kaplan, Omar A. Abdul-Rahman, Lisa M. Vincent, Amber Love, Khadija Belhassan, Karim Ouldim, Ihssane El Bouchikhi, Anju Shukla, Katta M. Girisha, Siddaramappa J. Patil, Nirmala D. Sirisena, Vajira H. W. Dissanayake, C. Sampath Paththinige, Rupesh Mishra, Eva Klein-Zighelboim, Bertha E. Gallardo Jugo, Miguel Chávez Pastor, Hugo H. Abarca-Barriga, Steven A. Skinner, Eloise J. Prijoles, Eben Badoe, Ashleigh D. Gill, Vorasuk Shotelersuk, Patroula Smpokou, Monisha S. Kisling, Carlos R. Ferreira, Leon Mutesa, Andre Megarbane, Antonie D. Kline, Amy Kimball, Emmy Okello, Peter Lwabi, Twalib Aliku, Emmanuel Tenywa, Nonglak Boonchooduang, Pranoot Tanpaiboon, Antonio Richieri-Costa, Ambroise Wonkam, Brian H. Y. Chung, Roger E. Stevenson, Marshall Summar, Kausik Mandal, Shubha R. Phadke, María G. Obregon, Marius G. Linguraru, and Maximilian Muenke. Noonan syndrome in diverse populations. *American Journal of Medical Genetics Part A*, 173 (9):2323–2334, 2017b. ISSN 1552-4833.

Paul Kruszka, Antonio R. Porras, Deise Helena de Souza, Angélica Moresco, Victoria Huckstadt, Ashleigh D. Gill, Alec P. Boyle, Tommy Hu, Yonit A. Addissie, Gary T. K. Mok, Cedrik Tekendo-Ngongang, Karen Fieggen, Eloise J. Prijoles, Pranoot Tanpaiboon, Engela Honey, Ho-Ming Luk, Ivan F. M. Lo, Meow-Keong Thong, Premala Muthukumarasamy, Kelly L. Jones, Khadija Belhassan, Karim Ouldim, Ihssane El Bouchikhi, Laila Bouguenouch, Anju Shukla, Katta M. Girisha, Nirmala D. Sirisena, Vajira H. W. Dissanayake, C. Sampath Paththinige, Rupesh Mishra, Monisha S. Kisling, Carlos R. Ferreira, María Beatriz de Herreros, Ni-Chung Lee, Saumya S. Jamuar, Angeline Lai, Ee Shien Tan, Jiin Ying Lim, Cham Breana Wen-Min, Neerja Gupta, Stephanie Lotz-Esquivel, Ramsés Badilla-Porras, Dalia Farouk Hussen, Mona O. El Ruby, Engy A. Ashaat, Siddaramappa J. Patil, Leah Dowsett, Alison Eaton, A. Micheil Innes, Vorasuk Shotelersuk, Ëben Badoe, Ambroise Wonkam, María Gabriela Obregon, Brian H. Y. Chung, Milana Trubnykova, Jorge La

Serna, Bertha Elena Gallardo Jugo, Miguel Chávez Pastor, Hugo Hernán Abarca Barriga, Andre Megarbane, Beth A. Kozel, Mieke M. van Haelst, Roger E. Stevenson, Marshall Summar, A. Adebowale Adeyemo, Colleen A. Morris, Danilo Moretti-Ferreira, Marius George Linguraru, and Maximilian Muenke. Williams–Beuren syndrome in diverse populations. *American Journal of Medical Genetics Part A*, 176(5):1128–1136, 2018. ISSN 1552-4833.

Stefanie Kung, Mark Walters, Peter Claes, Peter LeSouef, Jack Goldblatt, Andrew Martin, Shanti Balasubramaniam, and Gareth Baynam. Monitoring of Therapy for Mucopolysaccharidosis Type I Using Dysmorphometric Facial Phenotypic Signatures. In Johannes Zschocke, Matthias Baumgartner, Eva Morava, Marc Patterson, Shamima Rahman, and Verena Peters, editors, *JIMD Reports, Volume 22*, JIMD Reports, pages 99–106. Springer Berlin Heidelberg, Berlin, Heidelberg, 2015. ISBN 978-3-662-47453-2.

Clete A. Kushida, Deborah A. Nichols, Rik Jadrnicek, Ric Miller, James K. Walsh, and Kara Griffin. Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. *Medical care*, 50(Suppl):S82–101, July 2012. ISSN 0025-7079.

Sebastian Köhler, Leigh Carmody, Nicole Vasilevsky, Julius O. B. Jacobsen, Daniel Danis, Jean-Philippe Gourdine, Michael Gargano, Nomi L. Harris, Nicolas Matentzoglu, Julie A. McMurry, David Osumi-Sutherland, Valentina Cipriani, James P. Balhoff, Tom Conlin, Hannah Blau, Gareth Baynam, Richard Palmer, Dylan Gratian, Hugh Dawkins, Michael Segal, Anna C. Jansen, Ahmed Muaz, Willie H. Chang, Jenna Bergerson, Stanley J. F. Laulederkind, Zafer Yüksel, Sergi Beltran, Alexandra F. Freeman, Panagiotis I. Sergouniotis, Daniel Durkin, Andrea L. Storm, Marc Hanauer, Michael Brudno, Susan M. Bello, Murat Sincan, Kayli Rageth, Matthew T. Wheeler, Renske Oegema, Halima Lourghi, Maria G. Della Rocca, Rachel Thompson, Francisco Castellanos, James Priest, Charlotte Cunningham-Rundles, Ayushi Hegde, Ruth C. Lovering, Catherine Hajek, Annie Olry, Luigi Notarangelo, Morgan Similuk, Xingmin A. Zhang, David Gómez-Andrés, Hanns Lochmüller, Hélène Dollfus, Sergio Rosenzweig, Shruti Marwaha, Ana Rath, Kathleen Sullivan, Cynthia Smith, Joshua D. Milner, Dorothée Leroux, Cornelius F. Boerkoel, Amy Klion, Melody C. Carter, Tudor Groza, Damian Smedley, Melissa A. Haendel, Chris Mungall, and Peter N. Robinson. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Research*, 47(D1):D1018–D1027, January 2019. ISSN 0305-1048.

Judith H. Langlois and Lori A. Roggman. Attractive Faces Are Only Average. *Psychological Science*, 1(2):115–121, March 1990. ISSN 0956-7976.

Judith H. Langlois, Lori A. Roggman, and Lisa Musselman. What Is Average and What Is Not Average About Attractive Faces? *Psychological Science*, 5(4):214–220, July 1994. ISSN 0956-7976.

A. Lewyllie, J. Roosenboom, K. Indencleef, P. Claes, A. Swillen, K. Devriendt, C. Carels, M. Cadenas De Llano-Pérula, G. Willems, G. Hens, and A. Verdonck. A Comprehensive Craniofacial Study of 22q11.2 Deletion Syndrome. *Journal of Dental Research*, 96(12):1386–1391, November 2017. ISSN 0022-0345.

Liu Liu, Jiani Hu, Shuo Zhang, and Weihong Deng. Extended Supervised Descent Method for Robust Face Alignment. In C. V. Jawahar and Shiguang Shan, editors,

Computer Vision - ACCV 2014 Workshops, volume 9010, pages 71–84. Springer International Publishing, Cham, 2015. ISBN 978-3-319-16633-9 978-3-319-16634-6. URL `http://link.springer.com/10.1007/978-3-319-16634-6_6`.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep Learning Face Attributes in the Wild. *arXiv:1411.7766 [cs]*, November 2014. arXiv: 1411.7766.

M. Long, Y. Cao, J. Wang, and M. I. Jordan. Learning Transferable Features with Deep Adaptation Networks. *ArXiv e-prints*, February 2015.

Hartmut S. Loos, Dagmar Wieczorek, Rolf P. Würtz, Christoph von der Malsburg, and Bernhard Horsthemke. Computer-based recognition of dysmorphic faces. *European Journal of Human Genetics*, 11(8):555–560, August 2003. ISSN 1476-5438.

David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004. ISSN 1573-1405.

Daniel G. MacArthur and Chris Tyler-Smith. Loss-of-function variants in the genomes of healthy humans. *Human Molecular Genetics*, 19(R2):R125–130, October 2010. ISSN 1460-2083.

Rodrigo Minetto, Mauricio Pamplona Segundo, and Sudeep Sarkar. Hydra: an Ensemble of Convolutional Neural Networks for Geospatial Land Classification. *arXiv:1802.03518 [cs]*, February 2018. arXiv: 1802.03518.

Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. Learnable PINs: Cross-Modal Embeddings for Person Identity. *arXiv:1805.00833 [cs]*, May 2018. arXiv: 1805.00833.

Christoffer Nellåker, Fowzan S. Alkuraya, Gareth Baynam, Raphael A. Bernier, Francois P. J. Bernier, Vanessa Boulanger, Michael Brudno, Han G. Brunner, Jill Clayton-Smith, Benjamin Cogné, Hugh J. S. Dawkins, Bert B. A. deVries, Sofia Douzgou, Tracy Dudding-Byth, Evan E. Eichler, Michael Ferlaino, Karen Fieggen, Helen V. Firth, David R. FitzPatrick, Dylan Gration, Tudor Groza, Melissa Haendel, Nina Hallowell, Ada Hamosh, Jayne Hehir-Kwa, Marc-Phillip Hitz, Mark Hughes, Usha Kini, Tjitske Kleefstra, R. Frank Kooy, Peter Krawitz, Sébastien Küry, Melissa Lees, Gholson J. Lyon, Stanislas Lyonnet, Julien L. Marcadier, Stephen Meyn, Veronika Moslerová, Juan M. Politei, Cathryn C. Poulton, F. Lucy Raymond, Margot R. F. Reijnders, Peter N. Robinson, Corrado Romano, Catherine M. Rose, David C. G. Sainsbury, Lyn Schofield, Vernon R. Sutton, Marek Turnovec, Anke Van Dijck, Hilde Van Esch, Andrew O. M. Wilkie, and The Minerva Consortium. Enabling Global Clinical Collaborations on Identifiable Patient Data: The Minerva Initiative. *Frontiers in Genetics*, 10, 2019. ISSN 1664-8021.

Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On Spectral Clustering: Analysis and an algorithm. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 849–856. MIT Press, 2002.

Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and Transferring Mid-level Image Representations Using Convolutional Neural Networks. pages 1717–1724. IEEE, June 2014. ISBN 978-1-4799-5118-5.

Omkar Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep Face Recognition. *BMVC*, 1(3):6, September 2015.

Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the Fisher kernel for large-scale image classification. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 6314 LNCS, pages 143–156, 2010. ISBN 364215560X.

Vinay Uday Prabhu, Dian Ang Yap, Alexander Sheng-Shin Wang, and John Whaley. Covering up bias in CelebA-like datasets with Markov blankets: A post-hoc cure for attribute prior avoidance. *ArXiv*, abs/1907.12917, 2019. arXiv: 1907.12917.

E. Raff and J. Sylvester. Gradient Reversal Against Discrimination. *ArXiv e-prints*, July 2018.

Ana Rath, Natacha Marpillat, and Julie Christ. Prevalence of rare diseases: Bibliographic data. In *Rare Diseases Collection*. Orphanet Report Series, 2014.

Margot R. F. Reijnders, Robert Janowski, Mohsan Alvi, James Self, Ton J. van Essen, Maaike Vreeburg, Rob P. W. Rouhl, Servi J. C. Stevens, Alexander P. A. Stegmann, Jolanda Schieving, Rolph Pfundt, Katinke van Dijk, Eric Smeets, Connie T. R. M. Stumpel, Levinus A. Bok, Jan Maarten Cobben, Marc Engelen, Sahar Mansour, Margot Whiteford, Kate E. Chandler, Sofia Douzgou, Nicola S. Cooper, Ene-Choo Tan, Roger Foo, Angeline H. M. Lai, Julia Rankin, Andrew Green, Tuula Lonnqvist, Pirjo Isohanni, Sheeley Williams, Ilene Ruhoy, Karen S. Carvalho, James J. Dowling, Dorit L. Lev, Katalin Sterbova, Petra Lassuthova, Jana Neupauerova, Jeff L. Waugh, Sotirios Keros, Jull Clayton-Smith, Sarah F. Smithson, Han G. Brunner, Ceciel van Hoeckel, Mel Anderson, Virginia E. Clowes, Victoria Mok Siu, Paulo Selber, Richard J. Leventer, Christoffer Nellaker, Dierk Niessing, David Hunt, and Diana Baralle. PURA Syndrome: clinical delineation and genotype-phenotype study in 32 individuals with review of published literature. *Journal of Medical Genetics*, 55(2):104–113, February 2018a. ISSN 0022-2593.

Margot R. F. Reijnders, Kerry A. Miller, Mohsan Alvi, Jacqueline A. C. Goos, Melissa M. Lees, Anna de Burca, Alex Henderson, Alison Kraus, Barbara Mikat, Bert B. A. de Vries, Bertrand Isidor, Bronwyn Kerr, Carlo Marcelis, Caroline Schluth-Bolard, Charu Deshpande, Claudia A. L. Ruivenkamp, Dagmar Wieczorek, Diana Baralle, Edward M. Blair, Hartmut Engels, Hermann-Josef Lüdecke, Jacqueline Eason, Gijs W. E. Santen, Jill Clayton-Smith, Kate Chandler, Katrina Tatton-Brown, Katelyn Payne, Katherine Helbig, Kelly Radtke, Kimberly M. Nugent, Kirsten Cremer, Tim M. Strom, Lynne M. Bird, Margje Sinnema, Maria Bitner-Glindzicz, Marieke F. van Dooren, Marielle Alders, Marije Koopmans, Lauren Brick, Mariya Kozenko, Megan L. Harline, Merel Klaassens, Michelle Steinraths, Nicola S. Cooper, Patrick Edery, Patrick Yap, Paulien A. Terhal, Peter J. van der Spek, Phillis Lakeman, Rachel L. Taylor, Rebecca O. Littlejohn, Rolph Pfundt, Saadet Mercimek-Andrews, Alexander P. A. Stegmann, Sarina G. Kant, Scott McLean, Shelagh Joss, Sigrid M. A. Swagemakers, Sofia Douzgou, Steven A. Wall, Sébastien Küry, Eduardo Calpena, Nils Koelling, Simon J. McGowan, Stephen R. F. Twigg, Irene M. J. Mathijssen, Christoffer Nellaker, Han G. Brunner, and Andrew O. M. Wilkie. De Novo and Inherited Loss-of-Function Variants in TLK2: Clinical and Genotype-Phenotype Evaluation of a

Distinct Neurodevelopmental Disorder. *The American Journal of Human Genetics*, 102 (6):1195–1203, June 2018b. ISSN 0002-9297.

Jennifer L. Rennels, P. Matthew Bronstad, and Judith H. Langlois. Are attractive men's faces masculine or feminine? The importance of type of facial stimuli. *Journal of Experimental Psychology: Human Perception and Performance*, 34(4):884–893, 2008. ISSN 1939-1277(Electronic),0096-1523(Print).

Gillian Rhodes, Alex Sumich, and Graham Byatt. Are Average Facial Configurations Attractive Only Because of Their Symmetry? *Psychological Science*, 10(1):52–58, 1999. ISSN 0956-7976.

David J. Robertson, Robin S. S. Kramer, and A. Mike Burton. Face Averages Enhance User Recognition for Smartphone Security. *PLOS ONE*, 10(3):e0119460, March 2015. ISSN 1932-6203.

Hans Hilger Ropers. Genetics of Early Onset Cognitive Impairment. *Annual Review of Genomics and Human Genetics*, 11(1):161–187, 2010.

Rasmus Rothe, Radu Timofte, and Luc Van Gool. DEX: Deep EXpectation of Apparent Age from a Single Image. pages 252–257. IEEE, December 2015. ISBN 978-1-4673-9711-7.

Jorge Sanchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek. Image classification with the fisher vector: Theory and practice. *International Journal of Computer Vision*, 105(3):222–245, 2013.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A Unified Embedding for Face Recognition and Clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, June 2015. arXiv: 1503.03832.

Matthew Schultz and Thorsten Joachims. Learning a Distance Metric from Relative Comparisons. In S. Thrun, L. K. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 41–48. MIT Press, 2004.

Janneke H M Schuurs-Hoeijmakers, Edwin C Oh, Lisenka E L M Vissers, Mariëlle E M Swinkels, Christian Gilissen, Michèl a Willemsen, Maureen Holvoet, Marloes Steehouwer, Joris a Veltman, Bert B a de Vries, Hans van Bokhoven, Arjan P M de Brouwer, Nicholas Katsanis, Koenraad Devriendt, and Han G Brunner. Recurrent de novo mutations in PACS1 cause defective cranial-neural-crest migration and define a recognizable intellectual-disability syndrome. *American journal of human genetics*, 91 (6):1122–7, dec 2012. ISSN 1537-6605.

Dennis Craig Seager, Chung How Kau, Jeryl D. English, Wael Tawfik, Harry I. Bussa, and Abou El Yazeed M. Ahmed. Facial Morphologies of an Adult Egyptian Population and an Adult Houstonian White Population Compared Using 3d Imaging. *The Angle Orthodontist*, 79(5):991–999, July 2009. ISSN 0003-3219.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

John R. Shaffer, Ekaterina Orlova, Myoung Keun Lee, Elizabeth J. Leslie, Zachary D. Raffensperger, Carrie L. Heike, Michael L. Cunningham, Jacqueline T. Hecht, Chung How Kau, Nichole L. Nidey, Lina M. Moreno, George L. Wehby, Jeffrey C. Murray, Cecelia A. Laurie, Cathy C. Laurie, Joanne Cole, Tracey Ferrara, Stephanie Santorico, Ophir Klein, Washington Mio, Eleanor Feingold, Benedikt Hallgrimsson, Richard A. Spritz, Mary L. Marazita, and Seth M. Weinberg. Genome-Wide Association Study Reveals Multiple Loci Influencing Normal Human Facial Morphology. *PLOS Genetics*, 12(8):e1006149, 2016. ISSN 1553-7404.

L. Shao, F. Zhu, and X. Li. Transfer Learning for Visual Categorization: A Survey. *IEEE Transactions on Neural Networks and Learning Systems*, 26(5):1019–1034, May 2015.

Amanda J. Sharkey, editor. *Combining Artificial Neural Nets: Ensemble and Modular Multi-Net Systems*. Springer-Verlag, Berlin, Heidelberg, 1st edition, 1999. ISBN 978-1-85233-004-0.

A. I. M. Shaweesh, J. G. Clement, C. D. L. Thomas, and A. Bankier. Construction and use of facial archetypes in anthropology and syndrome diagnosis. *Forensic Science International*, 159 Suppl 1:S175–185, May 2006. ISSN 0379-0738.

Eamonn Sheridan, John Wright, Neil Small, Peter C Corry, Sam Oddie, Catherine Whibley, Emily S Petherick, Teena Malik, Nicole Pawson, Patricia A McKinney, and Roger C Parslow. Risk factors for congenital anomaly in a multiethnic birth cohort: an analysis of the Born in Bradford study. *The Lancet*, 382(9901):1350–1359, October 2013. ISSN 0140-6736.

P. Shukla, T. Gupta, A. Saini, P. Singh, and R. Balasubramanian. A Deep Learning Frame-Work for Recognizing Developmental Disorders. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 705–714, March 2017.

K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *ArXiv e-prints*, September 2014.

Karen Simonyan, Omkar Parkhi, Andrea Vedaldi, and Andrew Zisserman. Fisher Vector Faces in the Wild. *Procedings of the British Machine Vision Conference 2013*, pages 8.1–8.11, 2013. ISSN 1090235X.

Jitendra Kumar Sinha, Shampa Ghosh, and Manchala Raghunath. Progeria: A rare genetic premature ageing disorder. *The Indian Journal of Medical Research*, 139(5): 667–674, May 2014. ISSN 0971-5916.

Sivic and Zisserman. Video Google: a text retrieval approach to object matching in videos. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 1470–1477 vol.2, Nice, France, 2003. IEEE. ISBN 978-0-7695-1950-0.

A. Slavin Ross, M. C. Hughes, and F. Doshi-Velez. Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations. *ArXiv e-prints*, March 2017.

Rhona Smith. Regulation (EC) No 764/2008 of the European Parliament and of the Council. In *Core EU Legislation*, pages 183–186. Macmillan Education UK, London, 2015. ISBN 978-1-137-54501-5 978-1-137-54482-7.

Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical Networks for Few-shot Learning. *arXiv:1703.05175 [cs, stat]*, March 2017. arXiv: 1703.05175.

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway Networks. *arXiv:1505.00387 [cs]*, May 2015. arXiv: 1505.00387.

Y. Sun, X. Wang, and X. Tang. Deep Convolutional Network Cascade for Facial Point Detection. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3476–3483, June 2013.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going Deeper with Convolutions. *arXiv:1409.4842 [cs]*, September 2014. arXiv: 1409.4842.

Tai Sing Lee. Image representation using 2d Gabor wavelets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(10):959–971, October 1996.

Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, Columbus, OH, USA, June 2014. IEEE. ISBN 978-1-4799-5118-5.

B. Tiddeman, M. Burt, and D. Perrett. Prototyping and transforming facial textures for perception research. *IEEE Computer Graphics and Applications*, 21(5):42–50, July 2001. ISSN 0272-1716.

T. Tommasi, N. Patricia, B. Caputo, and T. Tuytelaars. A Deeper Look at Dataset Bias. *ArXiv e-prints*, May 2015.

Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *in CVPR*, 2011.

E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. Adversarial Discriminative Domain Adaptation. *ArXiv e-prints*, February 2017.

Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *Computer Vision (ICCV), 2015 IEEE International Conference on*, pages 4068–4076. IEEE, 2015.

A. Vedaldi and K. Lenc. Matconvnet – convolutional neural networks for matlab. In *Proceeding of the ACM Int. Conf. on Multimedia*, 2015.

Paul Viola and Michael J. Jones. Robust real-time face detection. *Int. J. Comput. Vision*, 57(2):137–154, May 2004. ISSN 0920-5691.

Tobias Vollmar, Baerbel Maus, Rolf P. Wurtz, Gabriele Gillessen-Kaesbach, Bernhard Horsthemke, Dagmar Wieczorek, and Stefan Boehringer. Impact of geometry and viewing angle on classification accuracy of 2d based analysis of dysmorphic faces. *European Journal of Medical Genetics*, 51(1):44–53, January 2008. ISSN 1769-7212.

Jacob A. S. Vorstman and Roel A. Ophoff. Genetic causes of developmental disorders. *Current Opinion in Neurology*, 26(2):128–136, April 2013. ISSN 1473-6551.

Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial Faces in-the-Wild: Reducing Racial Bias by Information Maximization Adaptation Network. *arXiv:1812.00194 [cs]*, December 2018a. arXiv: 1812.00194.

Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations. 2018b. arXiv: 1811.08489.

Xuehan Xiong and Fernando De la Torre. Supervised Descent Method and Its Applications to Face Alignment. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 532–539, jun 2013.

Xuehan Xiong and Fernando De la Torre. Supervised Descent Method for Solving Nonlinear Least Squares Problems in Computer Vision. *arXiv:1405.0601 [cs]*, May 2014. arXiv: 1405.0601.

Xuehan Xiong and Fernando De la Torre. Global supervised descent method. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2664–2673, Boston, MA, USA, June 2015. IEEE. ISBN 978-1-4673-6964-0. URL `http://ieeexplore.ieee.org/document/7298882/`.

Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. Learning Face Representation from Scratch. *arXiv:1411.7923 [cs]*, November 2014. arXiv: 1411.7923.

Xi Yin and Xiaoming Liu. Multi-Task Convolutional Neural Network for Pose-Invariant Face Recognition. *IEEE Transactions on Image Processing*, 27(2):964–975, February 2018. ISSN 1057-7149, 1941-0042. arXiv: 1702.04710.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond Empirical Risk Minimization. *arXiv:1710.09412 [cs, stat]*, October 2017. arXiv: 1710.09412.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

Qian Zhao, Kazunori Okada, Kenneth Rosenbaum, Lindsay Kehoe, Dina J. Zand, Raymond Sze, Marshall Summar, and Marius George Linguraru. Digital facial dysmorphology for genetic screening: Hierarchical constrained local model using ICA. *Medical Image Analysis*, 18(5):699–710, July 2014. ISSN 1361-8415.